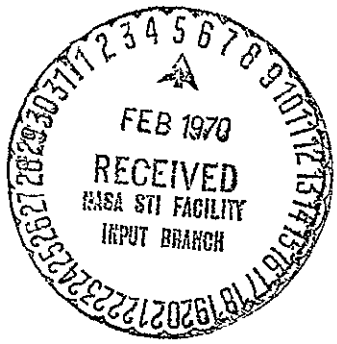


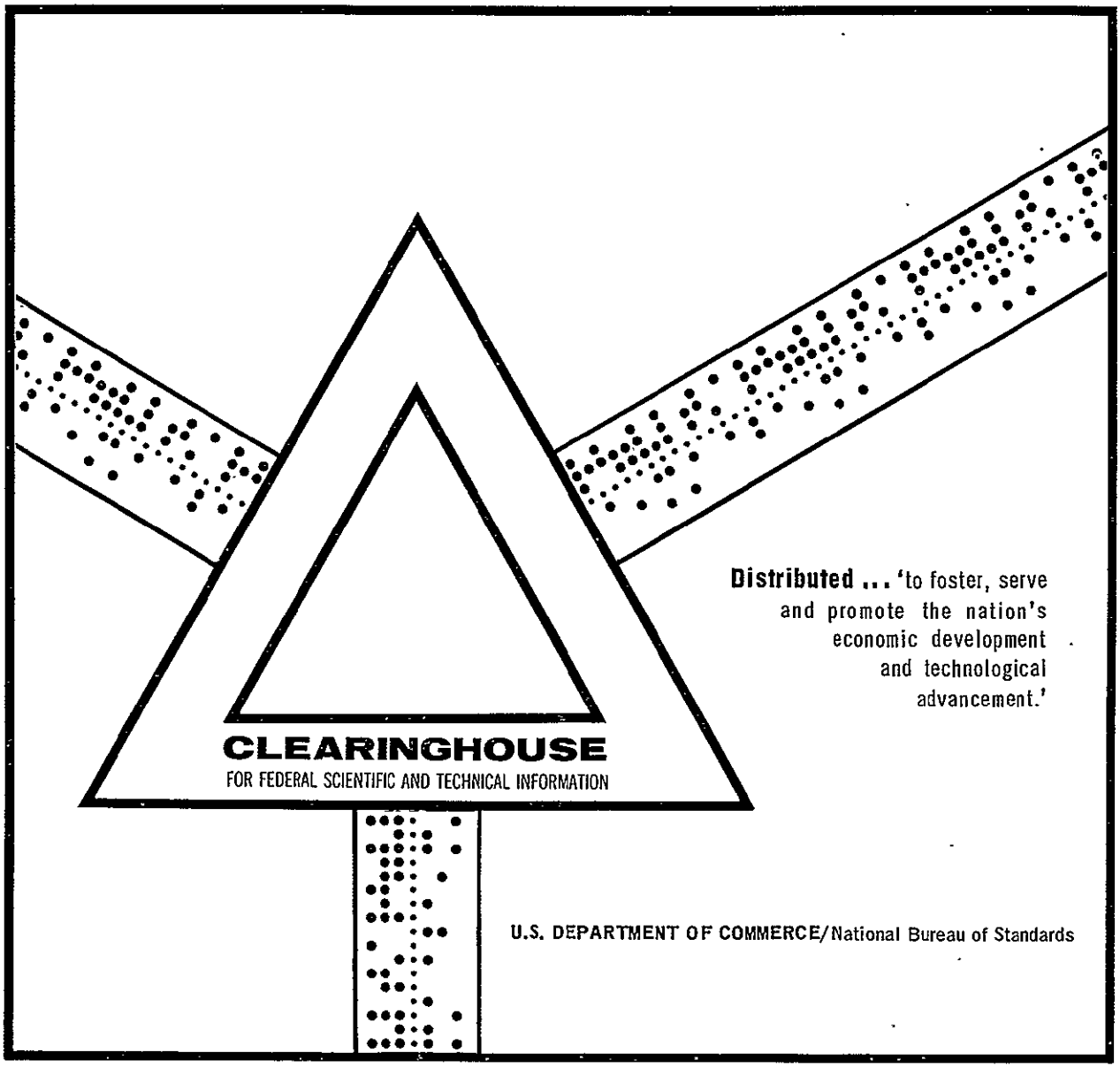
19

IDENTIFICATION OF LINEAR SYSTEMS

Thomas S. Englar

The Mathematical Sciences Group, Incorporated
College Park, Maryland

January 1970



RESEARCH REPORT

on

IDENTIFICATION OF LINEAR SYSTEMS

Prepared by

Thomas S. Englar

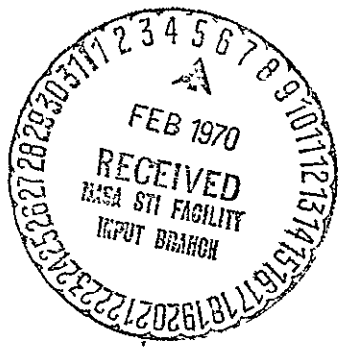
of
The Mathematical Sciences Group, Inc.
7100 Baltimore Avenue
College Park, Maryland

January 1970

Under Contract NAS 12-583

for

OCTA
NASA/Electronics Research Center
Cambridge, Massachusetts



PREFACE

This report has been written in partial fulfillment of Contract NAS 12-583 carried out by The Mathematical Sciences Group with the support of OCTA at NASA's Electronics Research Center.

The goal of work performed under this contract is the production of a digital computer program capable of identifying the dynamic characteristics of a human operator from knowledge of input-output data.

The component programs have been written and are documented herein. A certain amount of analysis has been done with experimental data and these results are also reported here.

We wish to take this opportunity to thank Dr. Richard Shirley for the assistance which he has rendered by his interest and suggestions.

TABLE OF CONTENTS

INTRODUCTION

This report describes how the identification problem has been approached in this work. Many of the details have been reported previously in Program Descriptions delivered to ERC. Parts of these are included as report appendices to provide complete information about all aspects of the analysis and computation. The body report concerns itself with the broader aspects of the system, referring to the appendices for deeper study.

Chapter I describes the context of linear systems analysis in which the problem is formulated, making specific our assumptions and the conditions the identified system must satisfy.

There are two basic subdivisions of the system. These are: obtaining the impulse response from the input and output functions and obtaining a realization from the impulse response by application of the B. L. Ho algorithm. Chapter II describes the mathematics involved in these two processes.

Chapter III describes how these methods are mechanized as computational techniques.

Chapter IV describes our preliminary results in system identification.

CHAPTER I

Background and Problem Statement

The ultimate goal of this work is the identification of the human operator in the sense that we wish to obtain a linear constant dynamical system which best approximates the human input-output behavior in a particular tracking task. In order to discuss the problem abstractly, we will assume that the system to be identified actually is a linear sationary dynamical system. By dynamical system we mean here a completely controllable, completely observable, finite dimensional system usually appearing as a set of differential equations relating the state $x(t)$ and the control $u(t)$ by

$$\dot{x} = Ax + Bu$$

and an algebraic relation relating the state and the output or vector of observables $y(t)$ by

$$y(t) = Cx(t)$$

Thus, our dynamical system can be represented by the triple of constant matrices $[C, A, B]$. As is well known, this representation is not unique since any similarity transformation S of A gives a new representation

$$[CS^{-1}, SAS^{-1}, SB].$$

We try to avoid this ambiguity by expressing the system in some canonical form.

This set of equations has the solution

$$y(t) = Ce^{tA}x(0) + \int_0^t Ce^{(t-\tau)A}Bu(\tau)d\tau \ .$$

Also characterizing this dynamical system is its impulse response

$$H(t) = Ce^{tA}B \ ,$$

or its transfer function

$$H(s) = C(sI-A)^{-1}B = \quad H(t) = \left[\frac{p_{ij}(s)}{q_{ij}(s)}\right]$$

The input-output relations sometimes appear in the form of an integral equation

$$y(t) = \int_0^t H(t-\tau)u(\tau)d\tau \ .$$

In all practical cases we define an additional variable $z(t)$ which is the state-dependent output $y(t)$ corrupted by some "noise" $v(t)$:

$$z(t) = y(t) + v(t) \ .$$

These remarks serve to delineate the context in which our problem is stated.

Problem 1: Given $\{z(t), u(t)\}$, defined on the interval $[0, T]$, obtain a minimal realization $[\hat{C}, \hat{A}, \hat{B}]$ such that

$$\int_0^T \left\| z(t) - \int_0^t \hat{C}e^{(t-\tau)\hat{A}}\hat{B}u(\tau)d\tau \right\|^2 dt$$

is minimal.

This problem does not consider the effect upon $z(t)$ of initial conditions on the state at time zero. Therefore, it will provide a good operating procedure only if $x(0)$ is zero. Unfortunately, it is impossible to place a human operator, such as a pilot, in zero-state condition. Furthermore, such a technique would limit applications of the program, since there are great advantages to examining some part of a long data run rather than only its initial phase. For instance, it allows the system, human or machine, to have a break-in or warmup period before taking data for analysis. Furthermore, one could wish to examine sequential data blocks in a long run to determine possible low frequency nonstationarity.

The most straightforward assumption which will enable such operations is:

Assumption: The system to be identified is asymptotically stable.

In addition we will proceed on the basis that the eigenvalues, initial conditions, and inputs are such that there exists a time $t_1 < T$ such that for computation purposes

$$y(t) = \int_0^t H(t-\tau)u(\tau)d\tau \quad \text{for} \quad t_1 \leq t \leq T \ .$$

We now state the problem to be solved.

Problem 2: Given functions $\{z(t), u(t)\}$ defined on the interval $[0, T]$, obtain a minimal realization $[\hat{C}, \hat{A}, \hat{B}]$ such that

$$\sigma^2 = \int_{t_1}^{T} \| z(t) - \int_{0}^{t} \hat{C} e^{\tau \hat{A}} \hat{B} u(t-\tau)\, d\tau \|^2 \, dt$$

is minimal.

The norm $\|\cdot\|$ used here is the usual Euclidean norm in finite dimensional space.

In what follows the mathematical methods, their numerical implementation, and recent numerical experiments will be described and analysed in some detail.

CHAPTER II

Mathematical Methodology

Involved in Problem 2 are two distinct subproblems, solutions to which we have programmed separately. The first is the definition of an approximating kernel $\hat{H}(t)$ such that σ^2 of Problem 2, is minimal.

The second is the definition of a system $[\hat{C}, \hat{A}, \hat{B}]$ such that, approximately,

$$\hat{H}(t) = \hat{C}e^{t\hat{A}}\hat{B} \ .$$

1. Obtaining $\hat{H}(t)$:

Without loss of generality, we restrict our discussion to scalar kernel functions $h(t)$.

The method used is basically a Rayleigh-Ritz procedure. However, important modifications in both the theory and the numerical techniques are implied by the fact that we are performing what, from an engineering viewpoint, might be called a second-level approximation problem. What is really desired is an approximation $\hat{h}(t)$ which minimizes

$$\varepsilon^2 = \int_0^{\infty} \|\hat{h}(t) - h(t)\|^2 dt \ .$$

But our problem constraints are such that we must be satisfied with solving Problem 2.

Problem 2 is mathematically equivalent (see Appendix I, sec. II-4), under the restrictions about stability which we have hypothesized, to minimizing

$$\int_0^{t_1}\int_0^{t_1}(\hat{h}(\tau) - h(\tau))Q(\tau,s)(\hat{h}(s) - h(s))ds d\tau = \|\hat{h} - h\|_Q^2 .$$

Here

$$Q(\tau,s) = \int_{t_1}^{T} u(t-\tau)u(t-s)dt$$

is a nonnegative definite symmetric kernel which is singular if u(t) is a band-limited function. That is, we are minimizing with respect to a pseudo-norm ($\|x\|_Q^2 = 0 \not\Rightarrow x = 0$).

If nothing else, the digital implementation which we use would serve to band-limit u(t) by the sampling theorem. Fortunately the singularity of Q does not seem to be a serious practical problem. The nonsingularity of Q is a measure of the amount of information about h(t) which is present in z(t). This is independent of additive output noise of course and our experiments with noise free data have indicated that the inputs now being used which are sums of ten to fourteen sinusoids are adequate for system determination.

We mention at this point -- it will be developed more fully later -- that proper selection of the input function u(t) can make Problem 2 mathematically equivalent to minimizing the error between the K^{th} fourier approximants of h(t) and $\hat{h}(t)$.

Returning therefore, to our Rayleigh-Ritz procedure for Problem 2, we assume that a set of functions $\{\ell_i(t)\}_o^K$ is available such that for each $h(t)$ of interest, there exists a linear combination

$$\hat{h}(t) = \sum_{k=o}^{K} \beta_k \ell_k(t) \tag{2.1}$$

with

$$\|\hat{h}(t) - h(t)\|$$

satisfactorily small.

This representation of $h(t)$ being decided upon, σ^2 is minimized with respect to the vector

$$\beta = \begin{bmatrix} \beta_o \\ \beta_1 \\ \vdots \\ \beta_K \end{bmatrix}$$

That is we compute

$$\hat{z}(t) = \int_0^t \hat{h}(\tau)u(t-\tau)d\tau = \sum_{k=o}^{K} \beta_k \int_0^t \ell_k(\tau)u(t-\tau)d\tau$$

and minimize

$$\int_{t_1}^{T} \|\hat{z}(t) - z(t)\|^2 dt \tag{2.2}$$

by manipulating β.

Defining a new set $\{f_i(t)\}_o^K$ of functions by

$$f_i(t) = \int_0^t \ell_i(\tau)u(t-\tau)d\tau$$

we find that the equation to be solved in the least square sense is

$$\sum_{k=o}^{K} \beta_k f_k(t) = z(t) \qquad t_1 \leq t \leq T \ .$$

Under very general conditions on $\{\ell_i\}_o^K$ and $u(\cdot)$ (Appendix I section II-3) we can show that $\{f_k(t)\}_o^K$ is a linearly independent set and there exists, therefore, a unique minimum of (2.2).

The result on linear independence cannot be stated briefly, but for

$$u(t) = \sum_{k=1}^{M} a_k \sin k\omega t, \qquad \prod_{k=1}^{M} a_k \neq 0$$

then, usually, the set $\{f_i\}_o^K$ will be linearly independent if

$$K + 1 \leq 2M$$

and $\{\ell_i\}_o^K$ is linearly independent.

A more detailed discussion of the effects of K and M is contained in Appendix I, section II-4. However, $M = 10$ is adequate for our purposes since we appear to obtain satisfactory approximation with $K = 16$.

There are two programs for obtaining $\hat{h}(t)$. RZ is used to compute examples -- that is sets of input-output data are computed internally and $\hat{h}$ then determined. XZ operates on externally generated

data. For instance the input-output functions for a human may be entered via punched cards.

2. Obtaining $[\hat{C}, \hat{A}, \hat{B}]$ from $\hat{h}(t)$:

The B. L. Ho method, [Reference 1], with some major modifications to alleviate the effects of noise, is used to obtain the system representation from the impulse response.

In section 1 of this chapter we dealt with single-input, single-output systems with no loss of generality. In computation of the representation, however, the multi-input, multi-output system $[\hat{h}_{ij}(t)]$ of impulse responses must be handled as a unit.

Probably the most important change to the Ho procedure which we have made is to work with the impulse response directly, obtaining a discrete system, then taking logarithms to obtain the continuous time representation.

A complete description of the Ho procedure as we have mechanized it in the Analysis program may be found in Appendix II, section II-1. The method for single-input, single-output systems will be outlined briefly here.

A sequence $\{\hat{a}_k\}$ is said to be of rank less than or equal to n if it can be generated from n-vectors c and b and an n^{th} order matrix A by the rule

$$a_k = cA^{k-1}b \ .$$

The B. L. Ho procedure takes a sequence (of finite rank), determines its rank, and exhibits the matrices $[c, A, b]$.

For $h(t) = ce^{tA}b$, the sequence

$$h_k = h((k-1)\delta) = ce^{(k-1)\delta A}b = c(e^{\delta A})^{k-1}b$$

is of finite rank and the Ho procedure will therefore give a discrete system similar to $[c, e^{\delta A}, b]$. This can then be transformed to a continuous system similar to $[c, A, b]$.

On the other hand, if we expand $h(t)$ in its Taylor Series

$$h(t) = \sum_{k=o}^{\infty} \frac{a_k}{k!} t^k$$

then

$$a_k = cA^k b$$

forms a sequence which satisfies the given condition and leads directly to a system similar to

$$[c, A, b] .$$

Generation of a_k from $\hat{h}(t)$ involves high order differentiation which is well known to be a poorly-conditioned operation on experimental data. In the past, both procedures were available; however, better results were obtained consistently with the sampled impulse response $\{h_k\}$ than with the Taylor Series coefficients $\{a_k\}$. For this reason the Analysis Program is now set up to handle only the discrete sequence $\{h_k\}$.

The program implementing the B. L. Ho procedure (MICARE) is described in Appendix II, the system logarithm program (CPC) is described in Appendix III. These two virtually complete the procedure; we have omitted the very simple routines describing how the sampled impulse response is obtained from the coefficients $\{\beta_i\}_0^K$. Input to MICARE is a sequence and if desired the Taylor coefficients $\{a_k\}$ could be entered, intermediate printout which now displays the discrete system would then give the continuous time representation.

REFERENCES

[1] B. L. Ho, "On Effective Construction of Realizations from Input-Output Descriptions," Ph.D. Dissertation, Stanford University (1966).

CHAPTER III

Implementation

The mathematics described in Chapter II is very straightforward and the implementation is very simple. The serious problems arise only in the presence of noise.

The first topic for consideration is: What should be the set of basic functions $\{\ell_i(t)\}_o^K$?

1. The Approximating Set

By our fundamental assumption, all $h(t)$ under consideration will decay to zero. It was felt therefore that the functions of the set $\{\ell_i\}$ should also satisfy this condition. This ruled out fourier approximation and the usual polynomial approximations.

Several sets of appropriate functions appear in the engineering literature (see W. H. Kautz, Transient Synthesis in the Time Domain, IRE Transactions-Circuit Theory, September 1954, pp. 29-39).

After considerable attention was given to these, especially to the laguerre functions, it was decided that none were satisfactory. At present we are using a set of exponentials $\{e^{\lambda_i t}\}$ whose eigenvalues lie in that region of the complex plane between $Re\lambda = -10$ and $Re\lambda = -0.1$.

Before describing the set of exponentials being used, a brief outline of the laguerre functions will be given which makes clearer the arguments against them.

The laguerre functions were chosen originally for two major reasons. They can be generated economically by using their recursion

relations, and their properties are very well-known, particularly the convergence of approximations using them (J. W. Head, Approximations to Transients by Means of Laguerre Series, Proc. Cambridge Philosophical Society, October 1956, pp. 640-651).

For arbitrary (real positive) p, the first few functions are:

$$\ell_o(t) = \sqrt{2p}\, e^{-pt}$$

$$\ell_1(t) = \sqrt{2p}\, e^{-pt}(2pt - 1)$$

$$\ell_2(t) = \sqrt{2p}\, e^{-pt}(2p^2t^2 - 4pt + 1)$$

$$\ell_3(t) = \sqrt{2p}\, e^{-pt}\left(\frac{4}{3} p^3t^3 - 6p^2t^2 + 6pt - 1\right)$$

$$\ell_4(t) = \sqrt{2p}\, e^{-pt}\left(\frac{2}{3} p^4t^4 - \frac{16}{3} p^3t^3 + 12p^2t^2 - 8p + 1\right)$$

$$\ell_5(t) = \sqrt{2p}\, e^{-pt}\left(\frac{4}{15} p^5t^5 - \frac{10}{3} p^4t^4 + \frac{40}{3} p^3t^3 - 20p^2t^2 + 10pt - 1\right).$$

In the frequency domain these functions have a particularly simple representation,

$$\ell_n(s) = \sqrt{2p}\, \frac{(p-s)^n}{(p+s)^{n+1}} \quad .$$

From this follows the interesting fact that the amplitude of the frequency response is independent of n, $|\ell_n(i\omega)| = |\ell_k(i\omega)|$.

The set satisfies the recursion formula

$$(n+1)\ell_n(t) = (2pt - 2n - 1)\ell_n(t) - n\ell_{n-1}(t) \quad .$$

The initial value is $\pm\sqrt{2p}$, $\ell_k(t)$ has k relative extrema of decreasing magnitude, and $\ell_k(t)$ for $p = 1$ is computationally zero at $2k + 7$. The most serious oscillations of ℓ_k occur near zero, where $\ell_k(t)$ behaves, to first order, like $e^{-(2k+1)pt}$. Table I shows the percentage error in Simpson's Rule integration of $e^{-\lambda t}$ for various numbers of integration intervals per time constant. (To avoid confusion here, by integration interval, we mean the interval between function evaluations, which is half of what is usually called the integration interval in Simpson's Rule.)

Assuming that we wish to integrate with a relative error of about 10^{-4}, we see that the integration interval δ must satisfy

$$\delta < \frac{1}{2.7p(2K+1)} .$$

In addition, to satisfy the decay property, $\ell_k(t) \approx 0$ for $t > t_1$ we must have $t_1 > 2K + 7$ for $p = 1$. Since p represents a linear time scaling, we may solve these relations for $p = 1$ and then modify the integration interval by a factor of $\frac{1}{p}$. This means that

$$\delta < \frac{1}{2.7(2K+1)} , \qquad t_1 > 2K + 7 .$$

In the computer program, the parameters determining t_1 are δ and INTST, the number of points omitted from fitting, by the relation

$$t_1 = (\text{INTST}-1)*\delta$$

Putting these together we find that

$$\frac{1}{2.7(2K+1)} \geq \delta \geq \frac{2K+7}{(\text{INTST}-1)} .$$

Solving this for K, and δ gives the following table

K	INTST	δ
0	19	.37
1	73	.123
2	150	.074
3	250	.0525
4	366	.041
5	510	.034
6	670	.028
7	856	.025
8	955	.022

At this point the hard facts of computer size intrude. We are at present limited to consideration of the function at 1600 points. It seems wasteful to devote less than half of these to the fitting interval.

In the light of all these factors, we chose

$K = 6$

$\delta = .025$

INTST = 800

as a working parameter set.

For completeness we must also ask if this integration interval is adequate to integrate the input satisfactorily.

For reasons which are explained in Appendix II, section II-4, the fundamental period appearing in the input should equal the fitting interval $T - t_1$. Therefore, the shortest period will be $\frac{T-t_1}{10}$ and have 80 points used for integration. The following table shows relative error in integrating sinusoids by Simpson's Rule, showing that we are easily within our desired error of 10^{-4}.

Intervals/period	Relative error
4	4.7%
8	.23%
12	$4.3 \cdot 10^{-4}$
16	$1.3 \cdot 10^{-4}$

The selection of parameters having been made, we must examine the systems which can be approximated satisfactorily. For this we refer to Head's paper, op.cit., to find that for arbitrary α and p,

$$e^{-\alpha t} = \frac{\sqrt{2p}}{\alpha+p} \sum_{k=o}^{\infty} \left(\frac{p-\alpha}{p+\alpha} \right)^k \ell_k(t)$$

Of course p, the eigenvalue of the laguerre functions is positive (or has positive real part) in our application, so this series is convergent if and only if α has positive real part, i.e., if our fundamental assumption of asymptotic stability is satisfied. However, we limit the series to seven terms; therefore, to satisfy our arbitrary desire for 10^{-4} relative error (approximately four significant digits) we must have

$$\left| \frac{\alpha-p}{\alpha+p} \right|^6 \simeq 10^{-4} .$$

This implies that

$$\left| \frac{\alpha-p}{\alpha+p} \right| \simeq 10^{-\frac{2}{3}} \simeq 0.215 .$$

The points α which satisfy

$$\left| \frac{\alpha-p}{\alpha+p} \right| = r$$

lie on a circle of radius

$$\frac{2r|p|}{1-r^2}$$

and center

$$p\,\frac{1+r^2}{1-r^2} \quad .$$

Unfortunately this does not cover nearly the desired area in the complex plane. For instance, in Figure 1, we show two circles to indicate the types of regions we could consider.

The preceding analysis has led us to an impasse which tells us that under the existing conditions we cannot approximate the desired spectrum of functions with a fixed set of laguerre functions. To illustrate, to encompass both $\alpha = 10$ and $\alpha = 0.1$, the best choice of p is 1 and the value of r will be $\frac{9}{11}$. In order to obtain 10^{-4} error, nearly 50 terms would be needed, requiring $\delta < .004$ and at the same time a fit inteval of 100 seconds (25000 points).

It was this problem that led to dropping use of the laguerre functions. The laguerre functions were used for a period while additive output noise was investigated, but for production use, the present system of distributed roots was adopted.

Several possibilities for modified use of laguerre functions suggested themselves. Figure 2 shows how four sets of laguerre functions could cover most of the desired region while staying within the computation constraints. Figure 3 shows an alternate configuration which while covering fewer oscillatory roots, blankets the real roots extremely well.

These choices of basis weaken the rationale for choosing laguerre functions. They use several sets which would require separate applications of the recursion formulas and complicate the analysis of convergence.

Therefore, in spite of the attractiveness of pioneering the use of laguerre functions with complex parameter p, this basis was abandoned.

Naturally the use of orthogonal exponentials was investigated, but there seemed to be no particular advantage for this application since the orthogonality which is needed is with respect to the kernel

$$Q(\tau,s) = \int_{t_1}^{T} u(t-\tau)u(t-s)dt$$

(see Chapter II, p. 9).

The set of eigenvalues now being used is:

$$\begin{aligned}
&-1.0\\
&-0.52\\
&-1.93\\
&-0.269\\
&-3.7\\
&-0.722 \pm i\ 0.25\\
&-1.39 \pm i\ 0.45\\
&-2.68 \pm i\ 0.85\\
&-0.374 \pm i\ 0.12\\
&-0.193 \pm i\ 0.08\\
&-0.2 \pm i\ 0.2
\end{aligned}$$

This set is not optimal, numerical experiments have shown that fitting error (normalized L_2 norm) varies widely on the real axis, ranging from 0.09 at -10.0 to -.02 at -0.1. The region in the complex plane with error norm less than 0.02 is shown in Figure 4.

The integral error on the real axis, for instance between -0.269 and -3.7 is much smaller than this, less than 0.007.

The present distribution is in geometric ratio on the real axis, with complex roots adjusted to give depth to the region at low damping. It is heavily weighted to fit well in the vicinity of -1.0. With no increase in the number of roots required a minimax distribution could be approximated to give a larger region analogous to that shown in Figure 4. In slightly less than full generality, the problem may be posed as follows:

Modified Distribution Problem: Given a probability density function P(s) defined on the left hand-plane and symmetric about the real axis, find n complex numbers $\{\lambda_i\}_1^n$ such that

$$E\left[\min_{\{a_i\}_1^n} \left\| e^{st} - \sum_{i=1}^{n} a_i e^{\lambda_i t} \right\|\right]$$

is minimized.

The roots listed above have been performing satisfactorily, giving much better approximation capabilities than the laguerre functions. The number of basis functions used is bounded above by twice the number of sinusoids in the input (See Appendix A). To insure good computational independence of the basis functions we keep the number of roots somewhat less than twenty when using a ten sinusoid input signal.

Integration accuracy has been examined for individual eigenfunctions. From Table I we see that there should be at least three integration intervals per time constant; we have seen previously that there should be at least twenty integration intervals per period. For

the present function and present input set, these constraints are satisfied when $\delta < 0.04$.

2. Integration Methods:

Trapezoidal integration was used initially but proved inaccurate. A procedure designed to convolve a tabulated function with laguerre functions was programmed and tested but was found to be no more accurate than trapezoidal integration because it required taking differences of large numbers. The integration now in use is Simpson's Rule.

Our computational object in the beginning was to be able to identify all eigenvalues with real parts between -10 and -0.1 and "reasonable" imaginary parts. To obtain satisfactory integration accuracy we should have an integration interval of about 0.03 and should have a total fit interval $[t_1,T]$ of length about 30.

The integration interval is compatible with that previously determined by the basis functions. The total fit interval of $800 \cdot 0.025 = 20$ seconds is less than the three time constants which would be ideal but does provide two time constants for the worst case (-0.1 eigenvalue).

3. Ho Procedure Modifications

Several modifications have been introduced into the standard multi-variable Ho procedure. The principal changes were designed to provide an overdetermined system, thus reducing noise effects, and to do this with the least additional computation cost.

To see most simply how this is done, consider a single-input, single-output system with sequence $\{a_k\}$. The usual procedure is to

form Hankel matrices H of increasing dimensions (2, 3, ···) until rank H no longer increases. For the low order systems which we usually consider, this means that only the first ten or 20 members of the sequence are used. When computing from exact data this is quite satisfactory. However, with noisy data it provides no averaging process.

For some time we were using Hankel matrices of order about twenty, thus using four or five times the data points actually required. Our experiments showed that we obtained better approximations with more data. Because the computation time was increasing with these large matrices, the following method was adopted.

Instead of using a Hankel matrix directly we used the first fifteen columns of the Hankel matrix. (The maximum system order allowed is fifteen.)

We take a large number of points, at present 98, and thus have a 98 by 15 matrix S composed of the leading columns of H. This method is possible only because the numerical procedure used to transform H works by orthogonalizing the column vectors of the matrix. Using such a program, we obtain the same system as obtained from the full Hankel matrix, but at considerable saving in computation.

The use of a large number of points gives a smoothing effect and in particular assures that lightly damped roots will appear to be stable. This is a serious problem. With short time spans in a noisy environment, lightly damped roots can easily appear unstable.

For multi-input, multi-output applications two initializing steps are required. The system is transposed if necessary so that there

are at least as many outputs p as inputs q. Each input is then assigned [15/q] columns and each output [98/p] rows. In each column, the elements associated with a given output component are grouped together in a block as in the usual Ho procedure. In each row the input number varies most quickly in order to provide more numerical significance. (It is computationally preferable to have independent columns coming first in the S matrix.)

These two changes, greatly overdetermined system and rectangular representation, are the principal distinguishing features of our mechanization of the B. L. Ho procedure.

While we get good smoothing in the computation of system poles, very little smoothing appears to take place in computation of the system numerator. Even in theory, there appears to be little that can be done about this problem once the fit has been established. For instance the zero-time response of the system to an impulse is given by one number, a_o, and that number must appear as the coefficient of the highest numerator power in the system transfer function.

Rank (system order) is controlled by a zero threshold determining linear independence and by IDERK, an input number specifying maximum rank allowed.

Single-input, single-output systems will be printed in companion form.

4. Nonlinear Kernel Functions

A very simple scheme for identifying nonlinear kernels has been mechanized in the RZ (self-generated) fit program. This assumes a

third order Volterra expansion with separable kernels whose basis elements are those used in the linear expansion.

Using the Volterra representation implies that the noise free output is

$$y(t) = \int_0^t h_1(t-\tau_1)u(\tau)d\tau_1 + \int_0^t \int_0^t h_2(t-\tau_1,\ t-\tau_2)u(\tau_1)u(\tau_2)d\tau_1 d\tau_2$$

$$+ \int_0^t \int_0^t \int_0^t h_3(t-\tau_1,\ t-\tau_2,\ t-\tau_3)u(\tau_1)u(\tau_2)u(\tau_3)d\tau_1 d\tau_2 d\tau_3 \ .$$

If the h_i are separable, then $y(t)$ has the representation

$$y(t) = y_1(t) + y_{21}(t)y_{22}(t) + y_{31}(t)y_{32}(t)y_{33}(t) \ ,$$

where each $y_{ij}(t)$ is the output of a linear system.

Such a representation can be used to approximate a large class of systems and is an exact model for systems having output which is the product of linear system outputs. The program has been checked on such systems.

By making the additional assumption that the linear kernels $h_{ij}(t)$ are drawn from the basis functions $\ell_i(t)$, mechanization of the nonlinear identification has been done with very little additional computation.

The problem of representation for such nonlinear systems is still open. The fit program gives only the coefficients of the expansion, a nonlinear analog of the B. L. Ho procedure is not available.

In addition to nonlinearities in the differential equations which can, at least in theory, be handled by the nonlinear kernels as

described above, there is another phenomenon appearing strongly in the human operator which cannot be represented in this way. This is the time delay caused by acquisition and processing time. This problem can be handled in the XZ Fit Program by shifting the system output with respect to the system input by an integer (LAG) number of integration intervals.

Number of intervals per time constant	Relative error
1.	5.0 e-3
1.1	3.5 e-3
1.2	2.5 e-3
1.3	1.8 e-3
1.4	1.4 e-3
1.5	1.0 e-3
1.6	8.1 e-4
1.7	6.4 e-4
1.8	5.1 e-4
1.9	4.1 e-4
2.	3.4 e-4
2.1	2.8 e-4
2.2	2.3 e-4
2.3	1.9 e-4
2.4	1.6 e-4
2.5	1.4 e-4
2.6	1.2 e-4
2.7	1.0 e-4
2.8	9.0 e-5
2.9	7.7 e-5
3.	6.8 e-5

TABLE I

Relative error in integrating $e^{-\lambda t}$ by Simpson's Rule.

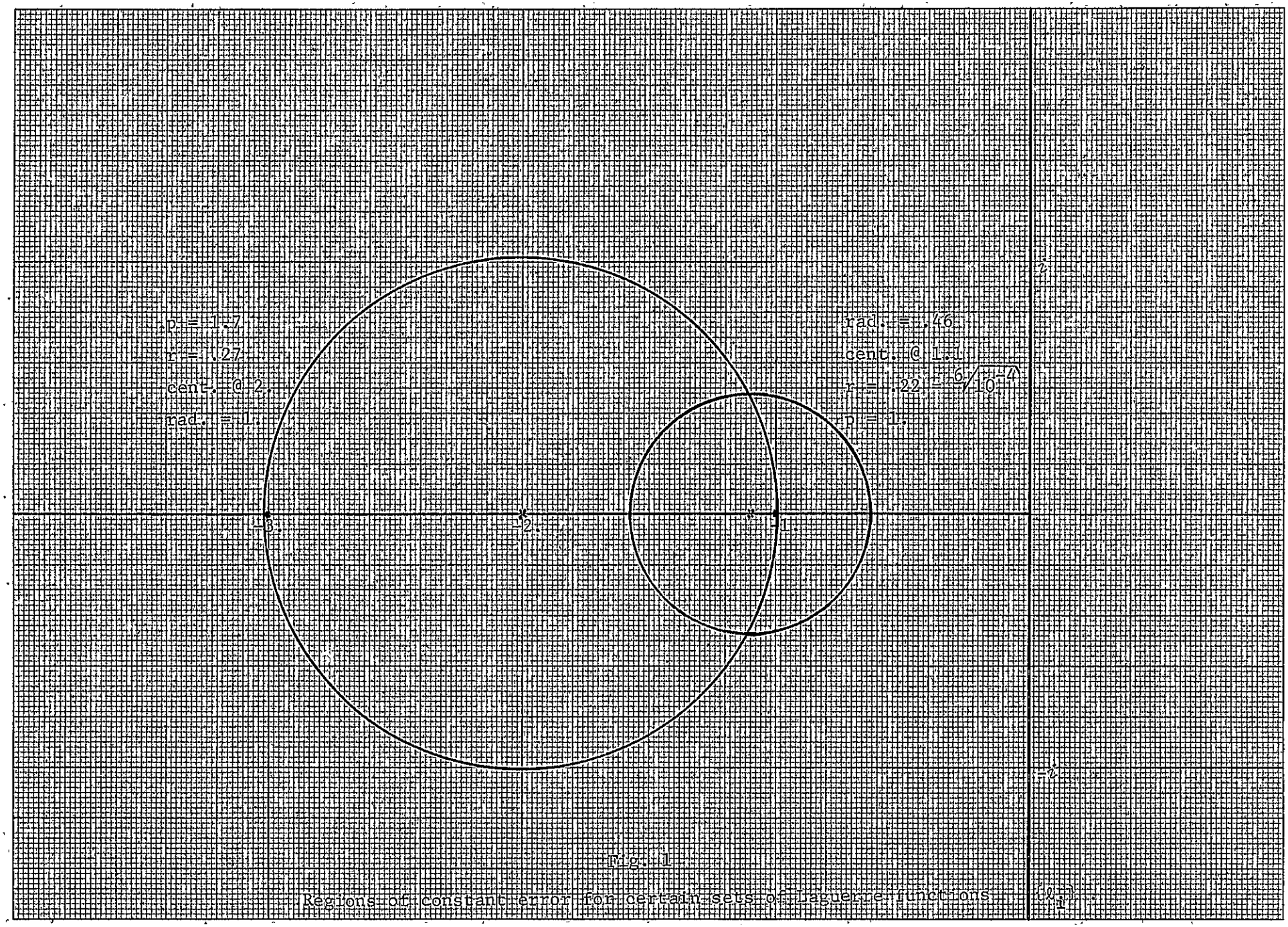


Fig. 1

Regions of constant error for certain sets of Laguerre functions

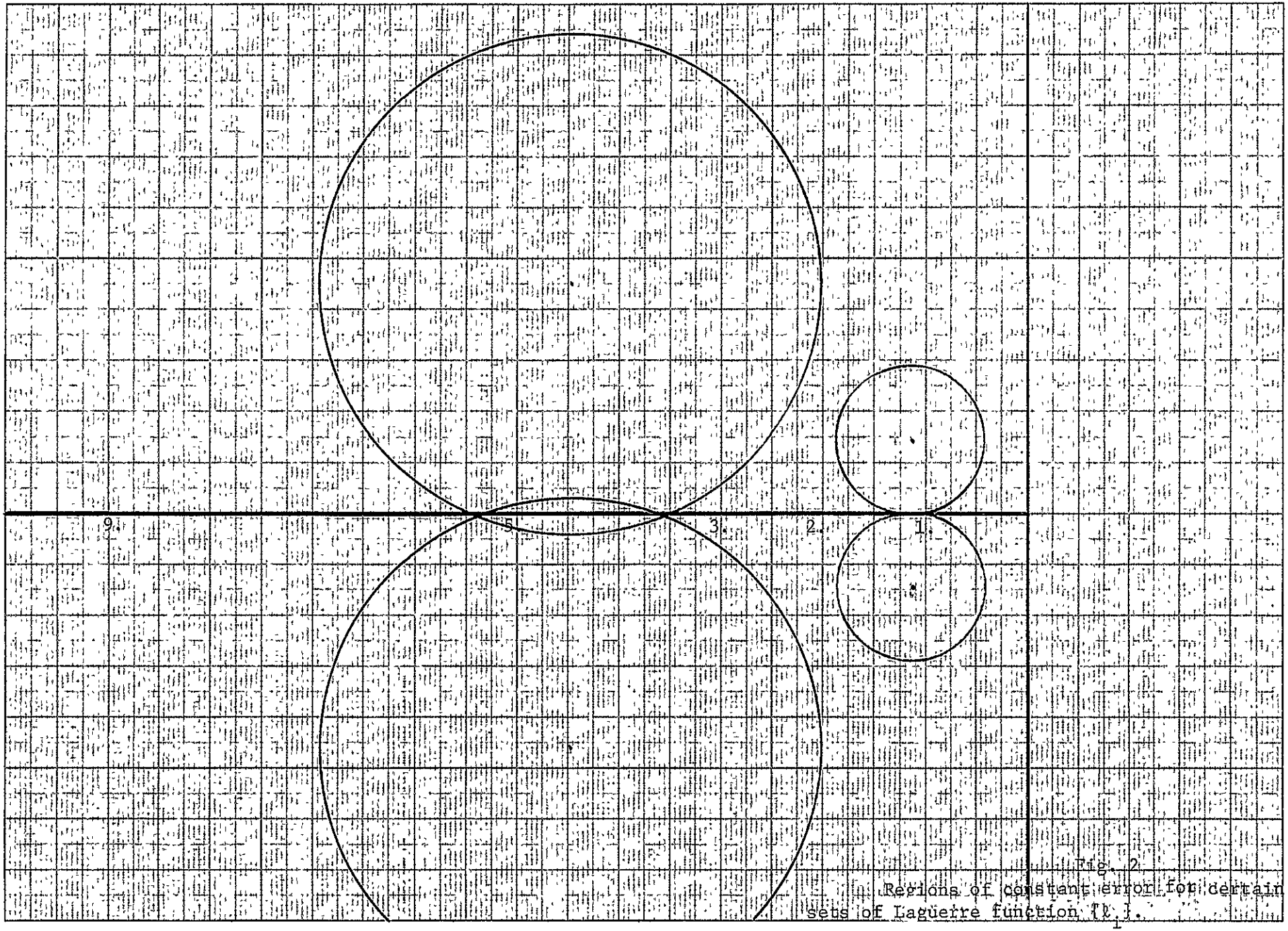


Fig. 2

Regions of constant error for certain sets of Laguerre function $\{l_j\}$.

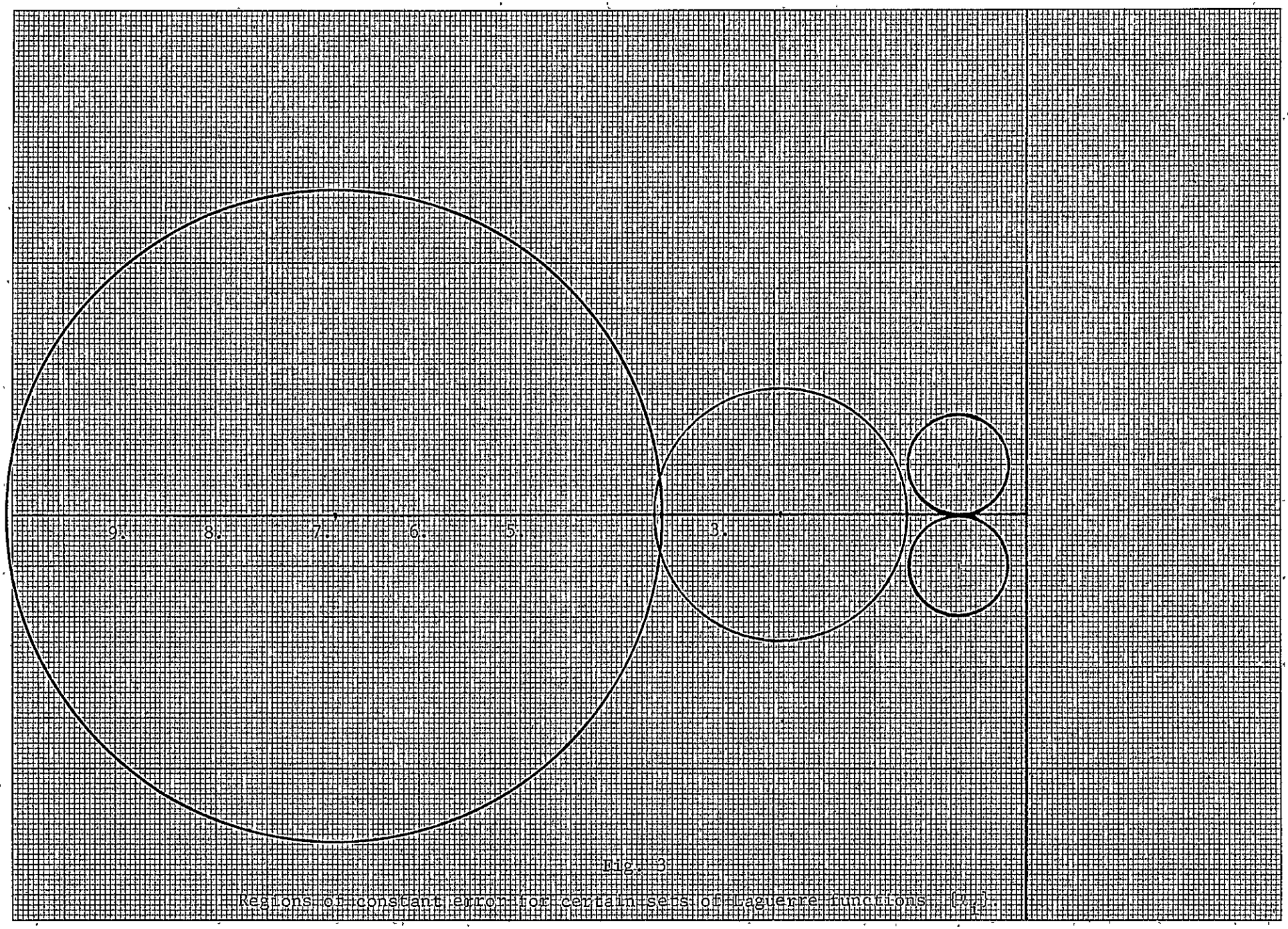
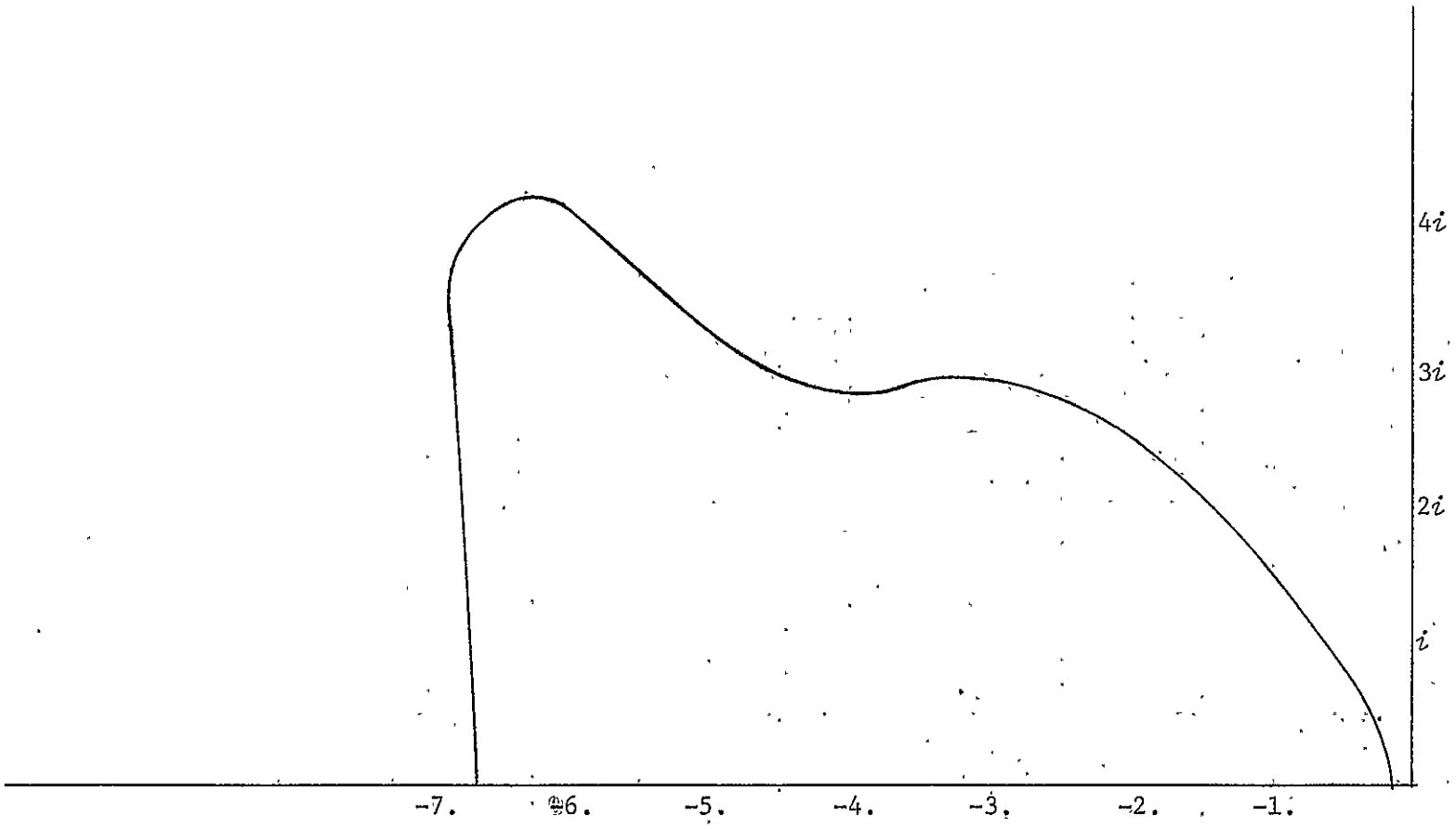


Fig. 3

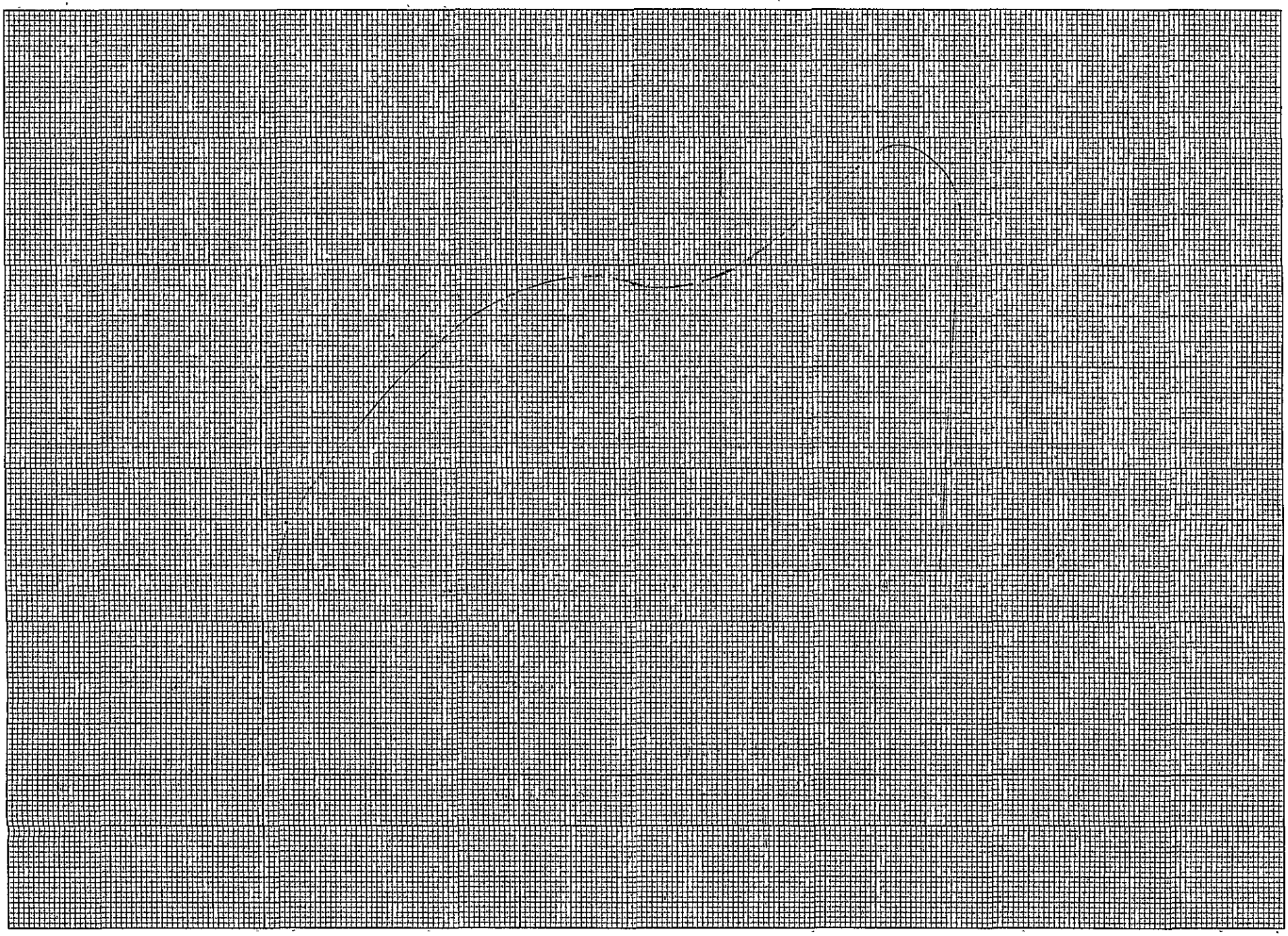
Regions of constant error for certain sets of Laguerre functions (P_n).



Region in which normalized error

$$\frac{\int_0^\infty [\hat{h}(t) - h(t)]^2 dt}{\int_0^\infty h^2(t) dt} \leq 0.0004$$

Fig. 4



CHAPTER IV

Numerical Results

The results reported here will give an idea of how well the system works on externally generated data. This is essentially noise-free operation with only the small errors introduced by the analog computer and recording devices affecting the accuracy. In addition two runs on human operator data are presented. Some brief comments concerning operation in the presence of noise and sensitivity of eigenvalues and system functions will be made.

System 1: This was a test run, using a system,

$$\frac{1}{(s+0.52)(s+1.93)}$$

which could be fitted exactly with

$$\beta_2 = 0.71, \quad \beta_3 = -0.71$$

and all others zero.

Taking the approximating impulse response obtained by averaging three sets of β's, each from a run of 884 points, we obtained a second order system (with a zero tolerance of 0.001) having transfer function

$$\frac{0.018s + 1.02}{s^2 + 2.49s + 1.03} = \frac{0.018s + 1.02}{(s + .523)(s + 1.97)}$$

This system has frequency response and impulse response varying only slightly from the correct values. The impulse response varies about

three per cent of peak near time zero. The phase and amplitude are less than 1% in error for frequencies below 4.36 rad/sec.

The startling aspect of these good results is that they were obtained with extremely poor values of the β vector. If self-generated data had been used the fit would have obtained $\beta_2 = -\beta_3 = 0.7$ very closely. The table below shows the leading β's for each of the three sets and their average.

	β_1	β_2	β_3
Run #1	-85.57	- 26.10	- 63.86
Run #2	2.98	-140.99	5.60
Run #3	498.04	502.74	-192.91
Average	138.48	111.88	- 83.72

This shows how very low noise levels translate into large deviations in the β vector. Fortunately these β errors do not always imply large errors in identification.

Computation of Taylor coefficients and their use in the B. L. Ho procedure has been dropped because of the inherent accuracy problems. To illustrate this we show the first few Taylor coefficients $\{s_k\}$, for this system -- exactly and as they were approximated.

h	s_k (exact)	s_k (fitted)
0	0	- 0.018
1	1	86.
2	-2.45	15.
3	5.	106.

This shows how very bad the point approximation (that is, the evaluation of the function and its derivatives) is at time t = zero, even when the approximant follows the function very closely. Indeed as we pointed out on page 25, the error at time zero is usually the largest error.

System 2: This system

$$\frac{88}{(s + 0.11)(s + 8.)}$$

was entered to see how well the system handled a large eigenvalue spread. Recall (see Chapter III) that 0.11 is near the boundary of the 0.02 error region and 8.0 is well beyond it. Actually the fitting error for 0.11 and 8.0 are, respectively, 0.013 and 0.045.

Taking the approximate system obtained by averaging three sets of 884 points, we obtained a second order system

$$\frac{-0.010s + 0.804}{s^2 + 7.12s + 0.896} = \frac{-0.010s + 0.804}{(s + 0.128)(s + 6.996)} .$$

This system very closely approximates the impulse response, the errors being greatest for large time because of the difference between 0.11 and 0.128. The frequency response is also quite satisfactory, amplitude ratio has less than 4% error below 4.36 rad/sec, but phase is not that good. Table 1 shows the frequency response, the columns are: input frequency in rad/sec; the amplitude ratio for the original system; the amplitude ratio obtained by the method of fourier coefficients from the experiment input and output functions; the amplitude ratio for the approximating system; and the three phase angles - exact, experimental, and approximate.

The responses are almost indistinguishable when graphed, except at the highest frequencies.

The root positions of the basis set are sufficient explanation for why the system roots pinched in to 7. and 0.13; there are no roots farther apart than 0.269 and 3.7.

The range of β for the three sets is again noteworthy.

β_1	β_2	β_3
-590.7	372.7	114.9
-155.4	5.2	74.1
170.6	32.3	-17.1
-191.8	136.7	57.3

System 3:

$$\frac{2}{s^2 + 2s + 2}$$

Taking the approximate impulse response obtained by averaging three sets of 884 points we obtained a second order system (when using a zero tolerance of 0.001) having transfer function

$$\frac{-0.029s + 2.12}{s^2 + 2.21s + 2.046} = \frac{-0.029s + 2.12}{(s + 1.10 \pm .091)}$$

This system has an excellent impulse response, differing from the exact values by less than 1% at peak and by 9% (less than 1% of peak) at the first minimum (.4 seconds). Frequency response also matched well showing a steady 4% error in amplitude and less than 2° in phase to 8.72 rad/sec. Table 2 presents the frequency response.

The three β vectors were analyzed individually and gave reasonable values, though not as good as the average, one set gave a third order system.

When the zero tolerance was reduced and the average data rerun, a third order system was obtained

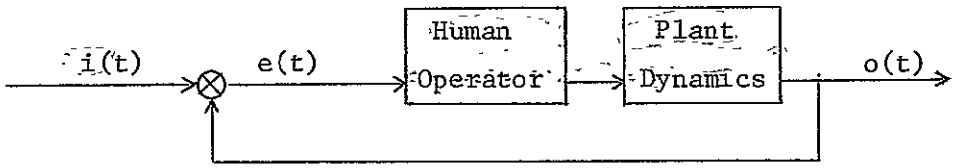
$$\frac{-0.029s^2 + 1.92s + 17.87}{s^3 + 11.14s^2 + 19.77s + 18.09}$$

$$= \frac{(s + 8.28)(-0.029s + 2.16)}{(s + 0.21)(s + 1.015 \pm .967)}$$

$$= \frac{s + 9.28}{s + 9.91} \quad \frac{-0.029s + 2.16}{s^2 + 1.982s + 1.97}$$

This system had an amplitude error of about 1% and slightly better phase than the other. The impulse response was essentially unchanged. Table 3 presents the frequency response.

System 4: The general block diagrom for human operator experiments is



Here, and also in System 5, the block marked "Human Operator" represents a trained pilot.

For the first run the plant dynamics were 1/s and we were attempting to fit the response o(t)/e(t). This violates the program rules somewhat since o(t)/e(t) has a pole at zero. However, the results were interesting.

Using the coefficients obtained from averaging over 11 sets of 884 points at 0.02 second intervals, the analysis program was applied with rank limitations of 2, 3, 4, and 5.

As the allowable rank was increased, the impulse response of the system obtained by the Ho procedure generally approximated more closely the impulse response as obtained from FIT. The frequency response varied greatly for the second, third, and fourth order systems, but the fourth and fifth order systems were the same to about 1%. This indicated that we could use the fourth order system as a good approximation.

The transfer function obtained is given in Figure 1 as is the frequency response from the experimental data and from the approximating transfer function.

System 5: Using the same allowable rank, the human operator test case was run again, this time with plant dynamics of $1/s^2$.

The transfer function we obtained is shown in Figure 2 as is the frequency response both from the experimental data and from the approximating transfer function.

Both of these fits compare well with results in the literature when we consider that these results are obtained from only a single run of 216 seconds.

Noise problems:

From the beginning, this approximation scheme has had difficulty in fitting the impulse response in the presence of noise. We have seen in System 1 that the β vector is wildly distorted merely by dealing with "clean" experimental data where signal to noise ratio is better than ten to one. Briefly, even small errors in output cause perturbations which the program attempts to fit.

The use of a prefilter was investigated but provided no better results in a noisy environment. This is to be expected. Since execution of the fitting process compares the output of the unknown system with the output of the basis systems to the same input, any additive function

which could not have originated by passing the given input through a linear filter will be excluded automatically.

ω	AR (exact)	AR (expt)	AR (FIT)	ϕ (exact)	ϕ (expt)	ϕ (FIT)
0.1164	0.687	0.702	0.664	- 47.5	- 48.1	-43.3
0.1745	0.533	0.539	0.531	- 59.0	- 60.7	-55.3
0.2909	0.353	0.354	0.361	- 71.4	- 71.4	-68.8
0.4363	0.244	0.245	0.252	- 79.0	- 80.3	-77.5
0.5818	0.185	0.185	0.192	- 83.5	- 83.7	-82.8
0.8727	0.124	0.124	0.129	- 89.0	- 90.5	-89.4
1.309	0.083	0.083	0.086	- 94.5	- 95.0	-96.0
1.745	0.061	0.062	0.064	- 98.7	-100.6	-101.1
2.618	0.040	0.040	0.041	-105.7	-106.9	-109.7
4.363	0.022	0.023	0.022	-117.2	-123.3	-123.5
6.545	0.013	0.013	0.013	-128.3	-129.5	-136.9
8.727	0.009	0.009	0.008	-136.8	-145.4	-146.9
15.71	0.003	0.003	0.003	-152.6	-160.7	-167.1
26.18	0.001	0.002	0.001	-162.8	-178.8	-183.6

Frequency Responses Associated with System $\dfrac{0.88}{s^2 + 8.11s + 0.88}$

TABLE 1

ω	AR (exact)	AR (expt)	AR (FIT)	ϕ (exact)	ϕ (expt)	ϕ (FIT)
0.1164	1.000	1.000	1.035	- 6.7	- 6.7	- 7.3
0.1745	1.000	1.000	1.033	-10.0	-10.1	-11.0
0.2909	0.999	0.999	1.027	-16.9	-17.1	-18.4
0.4363	0.995	0.998	1.014	-25.8	-26.0	-27.8
0.5818	0.986	0.985	0.992	-35.0	-35.3	-37.4
0.8727	0.935	0.934	0.915	-54.6	-55.1	-57.0
1.309	0.759	0.761	0.728	-83.8	-84.4	-84.5
1.745	0.549	0.550	0.532	-106.7	-107.6	-105.9
2.618	0.280	0.281	0.282	-132.8	-134.3	-131.8
4.363	0.104	0.102	0.109	-152.9	-154.8	-153.8
6.545	0.047	0.048	0.049	-162.2	-163.6	-165.6
8.727	0.026	0.028	0.028	-166.8	-171.5	-172.2
15.71	0.008	0.007	0.009	-172.7	-179.9	-184.1
26.18	0.003	0.003	0.003	-175.6	-191.1	-194.9

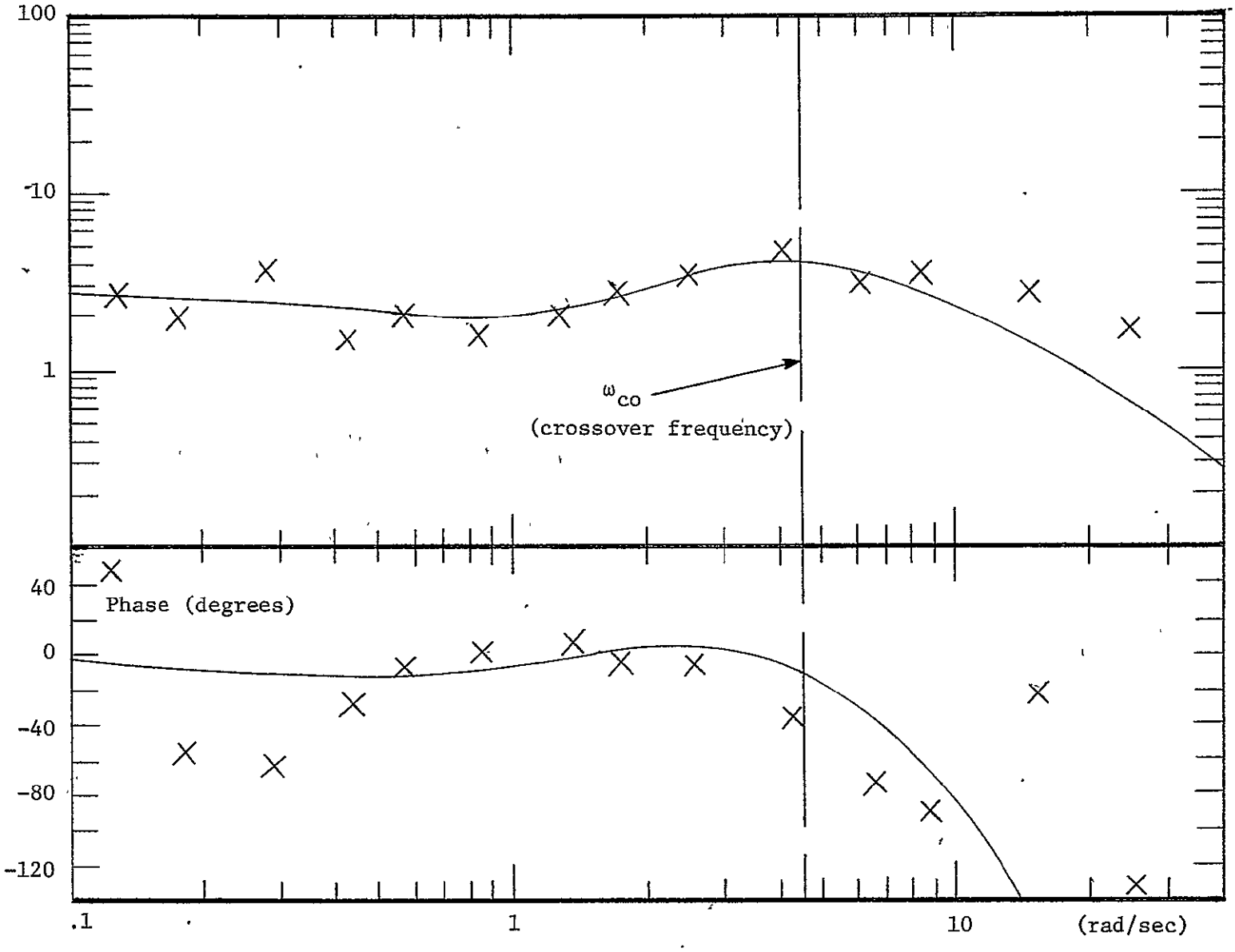
Frequency Responses Associated with System $\dfrac{2}{s^2 + 2s + 2}$

TABLE 2

ω	AR (exact)	AR (expt)	AR (FIT)	ϕ (exact)	ϕ (expt)	ϕ (FIT)
0.1164	1.000	1.000	.988	- 6.7	-6.7	- 6.6
0.1745	1.000	1.000	.989	-10.0	-10.1	- 9.9
0.2909	0.999	0.999	.989	-16.9	-17.1	-16.7
0.4363	0.995	0.998	.988	-25.8	-26.0	-25.5
0.5818	0.986	0.985	.982	-35.0	-35.3	-34.7
0.8727	0.935	0.934	.937	-54.6	-55.1	-54.6
1.309	0.759	0.761	.765	-83.8	-84.4	-84.4
1.745	0.549	0.550	.550	-106.7	-107.6	-107.9
2.618	0.280	0.281	.278	-132.8	-134.3	-134.3
4.363	0.104	0.102	.104	-152.9	-154.8	-154.6
6.545	0.047	0.048	.047	-162.2	-163.6	-164.9
8.727	0.026	0.028	.027	-166.8	-171.5	-170.8
15.71	0.008	0.007	.009	-172.7	-179.9	-182.2
26.18	0.003	0.003	.003	-175.6	-191.1	-193.3

Frequency Responses Associated with System $\dfrac{2}{s^2 + 2s + 2}$

TABLE 3



KEY:

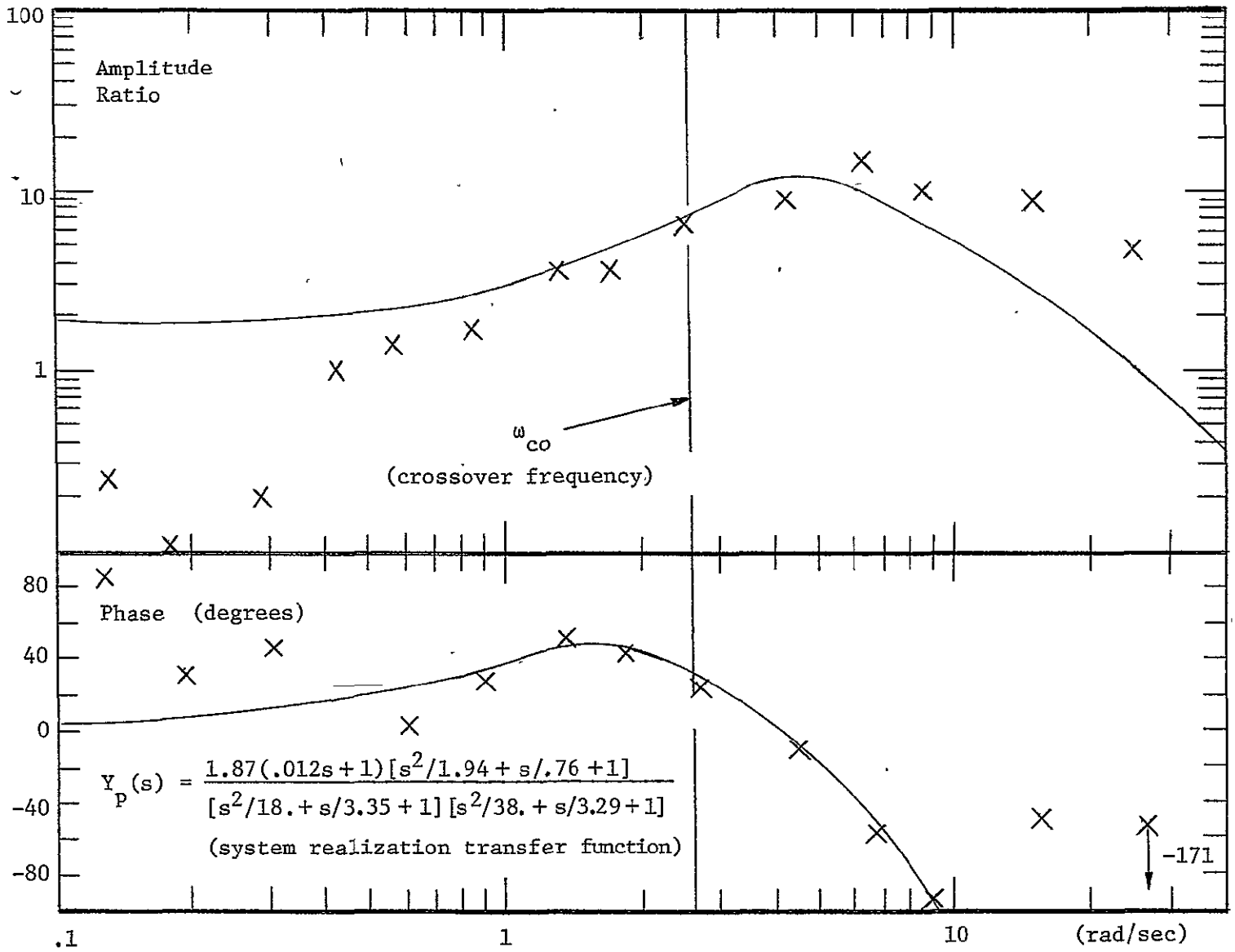
X measured using method of Fourier coefficients

——— results of B. L. Ho system realization, tol1 = .001

transfer function: $$Y_p(s) = \frac{2.59(-.0366s+1)[s^2/1.73 + s/.91 + 1]}{[\frac{s^2}{27.3} + \frac{s}{3.61} + 1](1.08s+1)(.14s+1)}$$

Fig. 1

FREQUENCY RESPONSE OF B. L. HO SYSTEM REALIZATION FOR HUMAN OPERATOR, $Y_c(s) = 1/s$



KEY:

× measured using method of Fourier coefficients

——— results of B. L. Ho system realization, tol 1 = .001

Fig. 2

FREQUENCY RESPONSE OF B. L. HO SYSTEM REALIZATION FOR HUMAN OPERATOR, $Y_c(s) = 1/s^2$

APPENDICES

These appendices contain functional descriptions of the programs and analysis of the operations which they perform. Procedural modifications such as transfer to the 360 are continuing at ERC. Therefore detailed program writeups were not included.

APPENDIX I

The Fit Program

I. Purpose

The purpose of this program is to generate the coefficients $\{\beta_i\}_0^K$ in a finite expansion

$$\sum_{i=0}^{K} \beta_i \ell_i(t) \tag{1}$$

for the impulse response of an asymptotically stable, linear, stationary dynamical system.

The data on which the program works is the input function $u(t)$ to the unknown system and the output $z(t)$ which is the system response corrupted by noise. Here $t\varepsilon[0,T]$.

The problem is solved by assuming the impulse response to be represented in the form (1).

This function then is convoluted with the input to produce an output which is a function of the finite vector

$$\beta = [\beta_0, \cdots, \beta_K]' .$$

This is compared with the actual output function z over a subinterval $[t_1,T]$ to allow the effect of initial conditions to decay and a least square solution obtained for β.

The actual mechanization works with discretized functions $\{u_i\}$ and $\{z_i\}$, $u_i = u((i-1)\delta)$.

The program described here works in a testing mode where the input and output sequences are generated internally from a known system. The deck described here uses a generalized inverse routine to solve for β. Other versions of the program, easily obtained from this one by modification, get the input-output sequences from externally generated cards and obtain β by inverting the normal matrix.

II. Mathematical Analysis

1. The Procedure.

Given the linear stationary dynamical system

$$\begin{aligned} \dot{x} &= Fx + Gu \\ y &= Hx \\ z &= y + v \ , \end{aligned} \tag{2.1}$$

where v is observational noise, we know that the output can be written as

$$z(t) = He^{tF}x(0) + \int_0^H He^{(t-\tau)F}Gu(\tau)d\tau + v \ .$$

By a change of variable, this can be rewritten as

$$z(t) = He^{tF}x(0) + \int_0^t He^{\tau F}Gu(t-\tau)d\tau + v \ .$$

From a knowledge of $z(t)$ and $u(t)$ only on some interval $[0,T]$, we want to obtain an estimate $\hat{h}(t)$ of

$$h(t) = He^{tF}G \ .$$

In order to do this, lacking knowledge of $x(0)$, we assume that F is asymptotically stable and that there exists a $t_1 < T$ such that in

$[t_1,T]$, $He^{tF}x(0)$ is very small compared with

$$\int_0^t He^{\tau F}Gu(t-\tau)d\tau \quad .$$

That is, we assume that on $[t_1,T]$,

$$z(t) = \int_0^t He^{\tau F}Gu(t-\tau)d\tau + v(t) \quad ,$$

and we then try to determine $\hat{h}(t)$ such that

$$\sigma^2 = \int_{t_1}^{T} [\int_0^t \hat{h}(\tau)u(t-\tau)d\tau - z(t)]^2 \, dt \tag{2.2}$$

is minimum.

Basically we use a Rayleigh-Ritz technique, that is we select a set of functions $\{\ell_i(t)\}$, which are "suitable" and represent $\hat{h}$ by linear combinations of the ℓ_i

$$\hat{h}(t) = \sum_{i=0}^{K} \beta_i \ell_i(t) \quad .$$

This reduces the problem to determining β such as to minimize σ^2.

$$\int_0^t \hat{h}(\tau)u(t-\tau)d\tau = \sum_{i=0}^{K} \beta_i \int_0^t \ell_i(\tau)u(t-\tau)d\tau \quad .$$

We call the integrals above new functions

$$f_i(t) = \int_0^t \ell_i(\tau)u(t-\tau)d\tau \quad . \tag{2.3}$$

Then it is the (nonorthogonal) basis set $f_i(t)$ upon which we will project $z(t)$ to determine β. We are fitting the function $z(\cdot)$ on $[t_1,T]$ with the expansion

$$\sum_{h=0}^{K} \beta_i f_i(\cdot) \quad .$$

Naturally we are interested in the linear independence of $\{f_i\}_o^K$. In addition we should determine whether or not the system (7.1) can be uniquely determined from a knowledge of only z and u. These two questions are intimately connected as the development in 3 will show.

Assuming the functions $\{f_i\}_o^K$ to be independent, however, we can proceed.

Rewriting 7.2 in terms of the $f_i(t)$ gives

$$\sigma^2 = \int_{t_1}^{T} [z(t) - \sum_{i=0} \beta_i f_i(t)]^2 \, dt, \qquad (2.2a)$$

which is then solved for the minimizing β vector.

2. Numerical Implementation.

A) The convolution integration in 2.3 is performed by Simpson's Rule, obtaining $f_i(t)$ at N+2 - INTST points on $[t_1,T]$. To expedite the mechanization, we insure an odd number of points on the interval $[0,t_1]$ by making INTST odd, and we make the number of points at which f_i is computed even by making N odd.

B) (2.2a) is minimized by using a generalized inverse routine to solve the linear finite system

$$[f_{ij}] \beta = z_i$$

where $f_{ij} = f_j(i\delta)$ and $z_i = z(i\delta)$.

3. Linear Independence of $\{f_i(t)\}_o^K$.

In order to investigate this we will consider only $u(t)$ of the type which we use, i.e.

$$u(t) = \sum_{k=1}^{M} c_k \sin \frac{k}{2} t , \qquad |c_k| = 1. \tag{2.4}$$

We further assume that all $\ell_i(t)$ are impulse responses of asymptotically stable, linear stationary dynamical systems; this is in fact a sine qua non for being "suitable" to our problem. Because we are looking only at steady-state output $z(t)$, $t \geq t_1$, after initial transients have subsided, the analysis is somewhat simpler. For any asymptotically stable system (2.1) the steady-state output $y(t)$ for input $\sin \frac{k}{2} t$ is

$$y(t) = A_k \sin \frac{k}{2} t + B_k \sin \frac{k}{2} t . \tag{2.5}$$

Since the $\ell_i(t)$ are impulse responses, $f_i(t)$ may be thought of as the output of a linear dynamical system to the input $u(t)$ and therefore is the sum of terms like (2.5).

Therefore we have

Lemma: A necessary condition for the functions $\{f_i(t)\}_o^K$ to be independent is that in (2.4), $M \geq \frac{K+1}{2}$.

Proof: $\{f_i(t)\}_o^K$ is a set of vectors from the $2M$ dimensional space spanned by

$$\{\sin \frac{k}{2} t, \quad \cos \frac{k}{2} t\}_1^M$$

therefore if $K + 1 > 2M$, the set is linearly dependent.

In fact, we can write the vector

$$f = \begin{bmatrix} f_o \\ f_1 \\ - \\ - \\ - \\ f_k \end{bmatrix}$$

as

$$f = Av \tag{2.4}$$

where

$$v = \begin{bmatrix} \sin \tfrac{1}{2} t \\ \cos \tfrac{1}{2} t \\ \sin \tfrac{1}{2} t \\ - \\ - \\ - \\ \cos \frac{M}{2} t \end{bmatrix}$$

and A is a constant matrix. Then $\{f_i\}_o^K$ is linearly dependent if there exists a constant vector $p \neq 0$ such that

$$p'f = 0 \quad .$$

Since A is $(K+1)$ by $2M$ it is clear that such a vector exists if $K + 1 > 2M$.

It is tempting to hypothesize that the $\{f_i\}_o^K$ are linearly independent if $M \geq \frac{K+1}{2}$ and the set $\{\ell_i\}_o^K$ is linearly independent. Unfortunately this is not true.

Counterexample:

$$\ell_o = e^{-\lambda t}$$

and

$$\ell_1 = \frac{(\eta-\lambda)(\mu^2+1)}{(\eta-\mu)(\lambda^2+1)} e^{-\mu t} + \frac{(\lambda-\mu)(\eta^2+1)}{(\eta-\mu)(\lambda^2+1)} e^{-\eta t}$$

have the same steady-state response to $\sin t$, i.e., $f_o(t) \approx f_1(t)$ for t large.

Since this implies that the systems

$$H = 1 \qquad F = -\lambda \qquad G = 1$$

and

$$H = [1,1] \qquad F = \begin{bmatrix} -\mu & 0 \\ 0 & -\eta \end{bmatrix} \qquad G = \begin{bmatrix} \frac{(\eta-\lambda)(\mu^2+1)}{(\eta-\mu)(\lambda^2+1)} \\ \frac{(\lambda-\mu)(\eta^2+1)}{(\eta-\mu)(\lambda^2+1)} \end{bmatrix}$$

have the same steady-state response to $u(t) = \sin t$, it is clear that we do have problems also in determining the system uniquely solely from input-output information.

Both questions can be answered easily however with the help of the following.

Corollary: Let $h(t)$ be the impulse response of a c.c. - c.o. , asymptotically stable linear stationary dynamical system. Let

$$\mathcal{L}\, h(t) = \frac{p(s)}{q(s)} \quad .$$

Then the steady-state response $f(t)$ of the system to $u(t)$, that is

$$f(t) \approx \int_0^t h(t-\tau)u(\tau)d\tau \qquad \text{for large } t \ ,$$

is zero if and only if $u(t)$ satisfies the homogeneous differential equation represented in the frequency domain by

$$p(s) \ ,$$

i.e., $\mathcal{L}^{-1}(p(s))\ u(t) = 0$.

Proof: This is a corollary to the much more general theorem by Leonard Weiss [1] .

Applying this to our case, we take the Laplace transforms of $\{\ell_i\}_o^K$, $\left\{\frac{p_i}{q_i}\right\}_o^K$ and compute, for all $\{a_i\}_o^K$

$$\frac{p(s)}{q(s)} = \sum_{i=0}^{K} a_i \ \frac{p_i}{q_i} \quad .$$

If deg $p(s) < 2M$ then the functions $\{f_i\}_o^K$ form a linearly independent set. In particular:

Case 1: The laguerre functions,

$$\frac{p_i(s)}{q_i(s)} = \frac{p_i(s)}{(s+1)^{i+1}}$$

Therefore $\deg p(s) < K + 1$, hence $2M \geq K + 1$ is both necessary and sufficient for linear independence.

Case 2: The Kautz function [2].

For the Kautz functions $\deg p \leq \deg p_K < K + 1$ for K odd and $\deg p = \deg p_{K+1} = K + 2$ for K even. In any case then, we have the same result, $2M \geq K + 1$ is both necessary and sufficient for linear independence.

Case 3: Arbitrary pole selection.

If we select

$$\ell_{2i} = e^{-\lambda_i(t)} \cos w_i t \qquad \text{for } i = 0,n;\ w_i \neq 0,$$

$$\ell_{2i+1} = e^{-\lambda_i(t)} \sin w_i t$$

and $$\ell_i = e^{-\lambda_i t} \qquad \text{for } i = 2n + 2,\ldots,K$$

with $w_i \neq w_j$ for $i \neq j$ and $\lambda_i \neq \lambda_j$ for $i \neq j$, $i, j \geq 2n + 2$ then $\deg p(s) < K + 1$. Again we have that $2M \geq K + 1$ is both necessary and sufficient for linear independence.

4. Uniqueness of Identification.

We wish to determine what system estimates

$$\hat{h}(t) = \sum_{i=0}^{K} \beta_i \ell_i(t)$$

can be obtained with fixed M and a set $\{\ell_i\}_0^K$, $K + 1 \leq 2M$, such that $\{f_i\}_0^K$ are linearly independent.

The counterexample in (3) can help our thinking about the problem. Letting $\lambda = 1$, $\mu = 2$, $\eta = 3$, we find that the systems

$$H_1 = 1 \qquad F_1 = -1 \qquad G_1 = 1$$

and

$$H_2 = [1,1] \qquad F_2 = \begin{bmatrix} -2 & 0 \\ 0 & -3 \end{bmatrix} \qquad G_2 = \begin{bmatrix} 5 \\ -5 \end{bmatrix}$$

have the same response to $u(t) = \sin t$. However they have impulse responses

$$h_1(t) = e^{-t}, \quad h_1(s) = \frac{1}{s+1}$$

and

$$h_2(t) = 5e^{-2t} - 5e^{-3t}, \quad h_2(s) = \frac{5}{(s+2)(s+3)} .$$

Figure 1 shows $h_1(t)$ and $h_2(t)$.

In their expansions in $\frac{1}{s}$ we have

$$h_1(s) \approx [1, -1, 1, -1, 1, \ldots]$$

$$h_2(s) \approx [0, 5, -25, 95, -325, \ldots] .$$

This shows that we can get an exact fit of the input-output relations and be very far wrong in the impulse response. We attempt to circumvent the problem by increasing M. For instance if in the previous example, we let $u(t) = \sin t + \sin 2t$, then we obtain the algebraic system

$$\begin{bmatrix} \frac{2}{5} & \frac{3}{10} \\ -\frac{1}{5} & -\frac{1}{10} \\ \frac{2}{8} & \frac{3}{13} \\ \frac{2}{8} & \frac{2}{13} \end{bmatrix} \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \\ \frac{1}{5} \\ -\frac{2}{5} \end{bmatrix}$$

Here β_o and β_1 are respectively the coefficients of the functions $\ell_o = e^{-2t}$ and $\ell_1 = e^{-3t}$ which will minimize (2.2). The optimal β_i are

$$\beta_o = 4.01572$$

$$\beta_1 = 3.62098$$

and the impulse response appears in Fig. 1.

The most unfortunate aspect of the procedure is that the error

$$\varepsilon^2 = \int_0^\infty (h(t) - \hat{h}(t))^2 dt$$

is not a monotonic function of σ^2, in 2.2, for fixed K. For instance in this case the vector which minimizes ε^2 is

$$\beta_o = 3\frac{1}{3}$$
$$\beta_1 = 2.5 \quad .$$

The impulse response for this fit appears in Fig. 1 also.

Note, in fact, that ε^2 does not necessarily decrease for fixed M as K increases. In fact we can obtain a better ε^2 fit with $\beta_o = \frac{3}{2}$, $\beta_1 = 0$, which is the minimum σ^2 fit for K = 0, M = 1 than by minimizing σ^2 for K = 1, M = 1.

Remark:

$$\int_0^\infty (e^{-t} - \beta_o e^{-2t} - \beta_1 e^{-3t})^2 dt$$
$$= \frac{15\beta_o^2 + 24\beta_o\beta_1 - 40\beta_o + 10\beta_1^2 + 30\beta_1 + 30}{60} \quad .$$

Now we see that there are two aspects of the uniqueness question. Let us take an asymptotically stable system (2.1) of order n and record its steady-state output for $2M \geq n$. Then there is only one system of order n which will give that output.

On the other hand if the eigenvalues are unknown and we use some arbitrary set of functions $\{\ell_i(t)\}_o^K$ then it is not necessarily true that we are fitting the impulse response more closely as K increases with M remaining fixed, even though the functions $\{f_i(t)\}_o^K$ are linearly independent. The example above shows this very clearly.

Moreover it is clear that when $K + 1 = 2M$, then σ^2 can be made zero while ϵ^2 remains large.

Before attempting any conclusions about uniqueness or operational procedures we should obtain a better idea of the mathematical principles which underlie the process we are using. That will be done for a slightly idealized variation in the development which follows. The idea may be stated easily. Instead of minimizing $||L\beta - h||$, where

$$L = [\ell_o(t),\ldots,\ell_K(t)],$$

we are minimizing $||L\beta - h||_Q$, where Q is a non-negative definite symmetric kernel.

What our program does is to minimize

$$\sigma^2 = \int_{t_1}^{t} [F\beta - z(t)]^2 dt ,$$

where F is the $K + 1$ - component new vector with

$$F_i = \int_0^t u(t-\tau)\ell_{i-1}(\tau)d\tau \quad .$$

Using the definition of F we can rewrite σ^2 as

$$\int_{t_1}^{T} \int_0^t [\beta L(\tau)-h(\tau)]u(t-\tau) \int_0^t u(t-s)[L(s)\beta-h(s)]ds d\tau dt .$$

We now assume explicitly that $L(t) = 0 = h(t)$ for $t \geq t_2$ and that $t_1 \geq t_2$. Interchanging integrals then gives us

$$\int_0^{t_2}\int_0^{t_2} [\beta L(\tau)-h(\tau)] \int_{t_1}^{T} u(t-\tau)u(t-s)dt\ [L(s)\beta-h(s)]ds d\tau$$

$$\int_0^{t_2}\int_0^{t_2} [\beta L(\tau)-h(\tau)]Q(\tau,s)[L(s)\beta-h(s)]ds d\tau$$

$Q(\tau,s)$ is clearly non-negative definite symmetric. Furthermore, with

$$u_M(t) = \sum_{k=1}^{M} \sin k\,\omega\, t \quad ,$$

if $T - t_1$ is a multiple of $\frac{2\pi}{\omega}$, then the components of $u(t)$ are orthogonal, and ε^2 associated with $u_{M+1}(t)$ is less than $u_M(t)$. (This follows from the fact that the eigenfunctions of $Q_M(s,t)$ with nonzero eigenvalues are orthogonal and coincide with a subset of the eigenfunctions of $Q_{M+1}(s,t)$.)

One thing that is not clear from this is the speed with which

$$||L\beta - h||_Q \to ||L\beta - h|| \quad .$$

Treated as a periodic function, those components of L which are nonzero at zero have a discontinuity at zero and therefore have considerable high frequency power. In fact because of this discontinuity, we cannot prove simply that

$$||L\beta - h||_Q \to ||L\beta - h||$$

and we cannot expect convergence better than $\frac{1}{M}$.

We can draw some recommendations from this analysis for use in our operational procedures.

1) The input should contain a constant.

2) The lowest frequency, ω, appearing in u should be such that

$$T - t_1 = \delta(N + 1 - INTST)$$

is a multiple of $\frac{2\pi}{\omega}$.

It is interesting that when

$$u = \frac{I}{4} + \sum_{k=1}^{M} L \sin k\omega t$$

then the procedure, in effect, takes the Mth order approximant $\hat{L}$ of L and minimizes $||\hat{L}\beta - h||$.

REFERENCES

[1] Leonard Weiss, "On a Question Related to the Control of Linear Systems," IEEE Transactions on Automatic Control vol. AC-9, Number 2, April 1964.

[2] B. L. Ho, "On Effective Construction of Realizations from Input-Output Descriptions," Ph.D. Dissertation, Stanford University (1966).

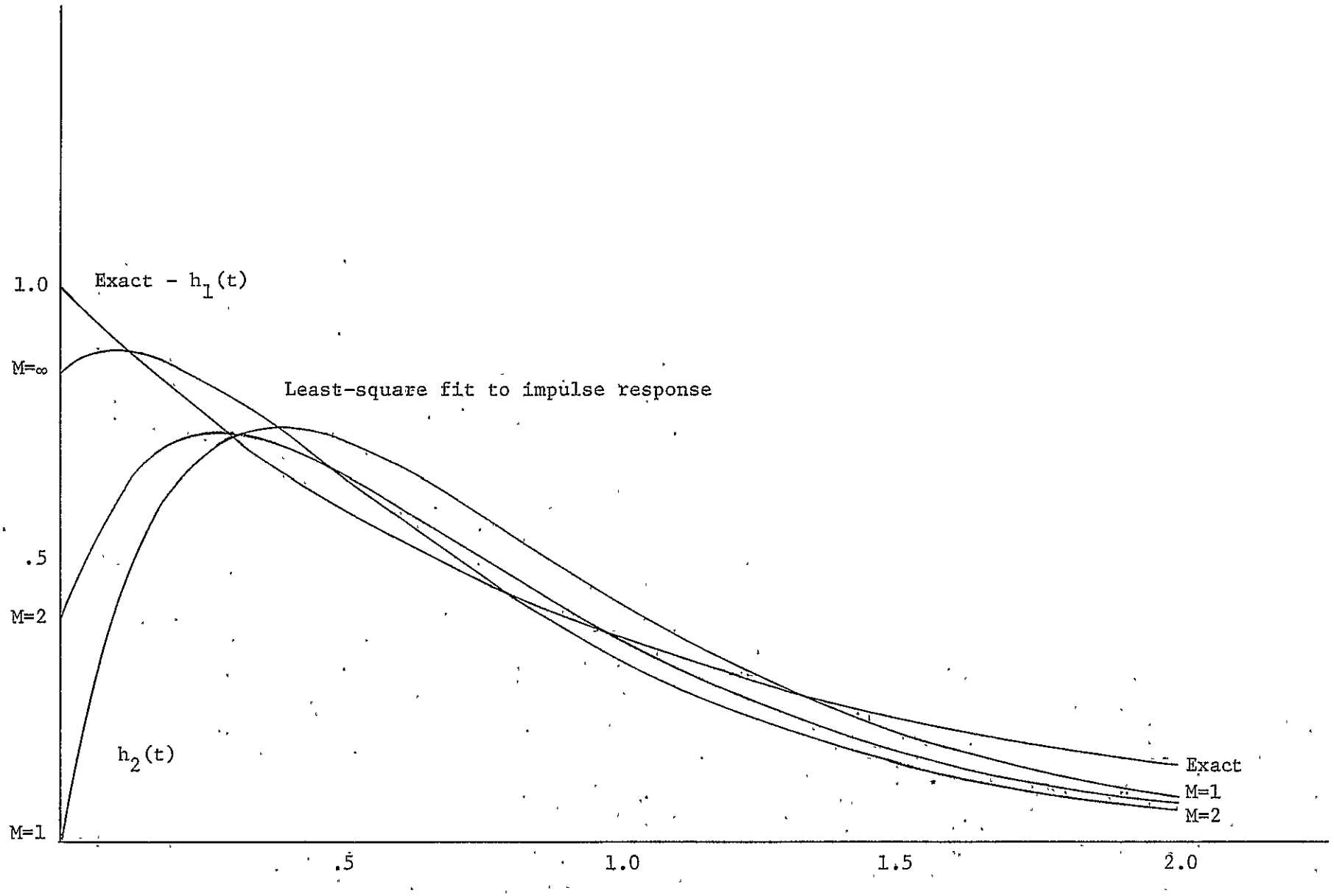
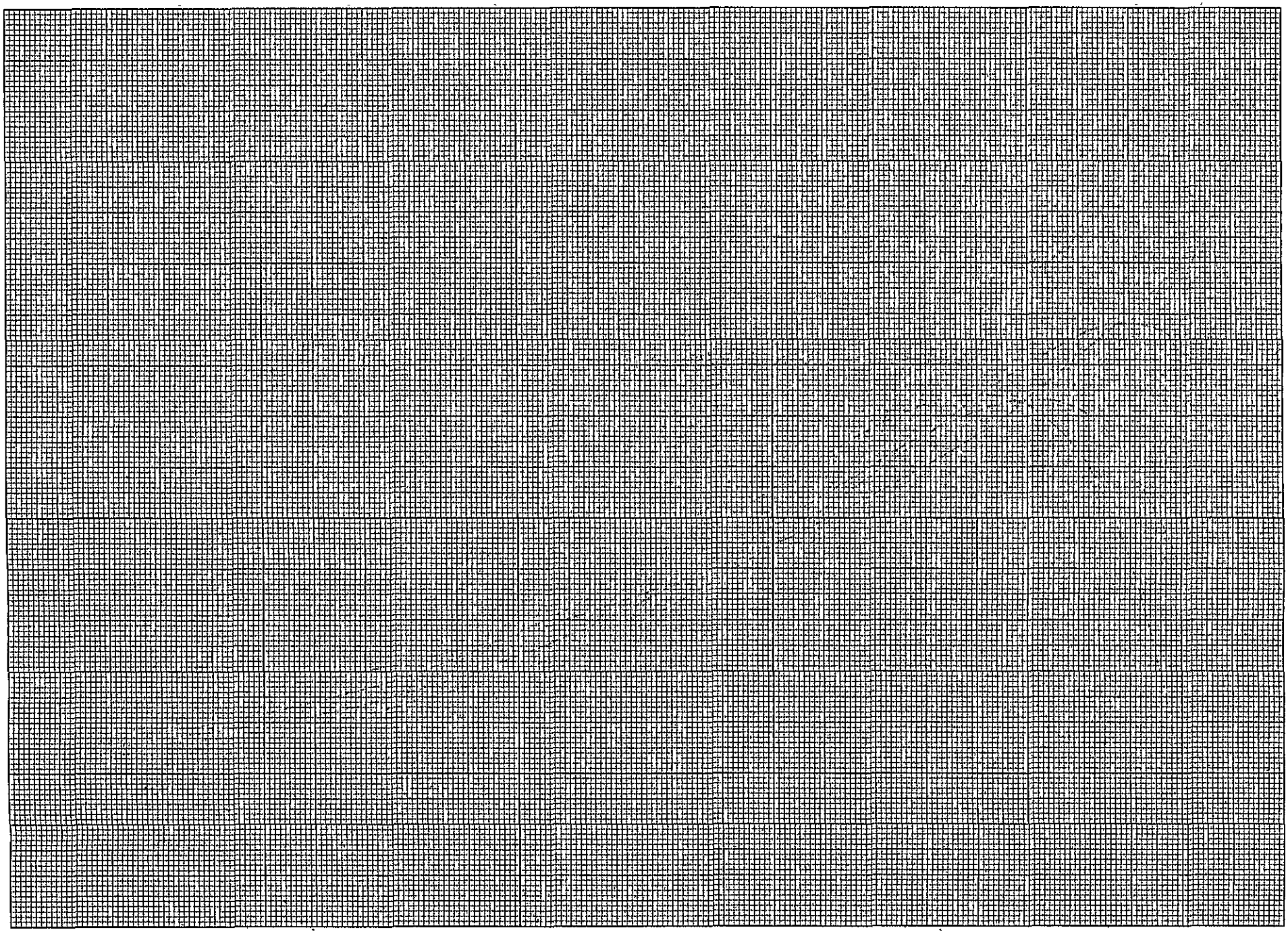

1.0
Exact - $h_1(t)$
M=∞
Least-square fit to impulse response
.5
M=2
$h_2(t)$
M=1
Exact
M=1
M=2
.5
1.0
1.5
2.0

Fig. 1



APPENDIX II

The Subroutine Micare

SUBROUTINE MICARE (SSUBR, N, TOLI, NST, H3, IDERK)

I. Purpose

We are given the N vector $S = \{s_k\}$ and wish to find an r-dimensional, constant linear dynamical system, $[c, \phi, \gamma]$ in companion form, with $c = [1, 0, \ldots, 0]$ such that, approximately,

$$c\phi^{k-1} = s_k \qquad k = 1, \ldots, N .$$

This is the primary task of subroutine MICARE - the implementation of the B.L. Ho procedure.

In addition, however, it calls subroutine CPC (see Appendix III) in order to obtain an r-dimensional, constant linear dynamical system $[c, A, b]$ in companion form, with $c = [1, 0, \ldots, 0]$ such that

$$c\, e^{k\sigma A}\, b = s_{k+1}, \qquad k = 0, \ldots, N-1.$$

Essentially CPC finds the continuous-time system $[c, A, b]$ from which the discrete system $[c, \phi, \gamma]$ arises. This is under the assumption that the input vector s is the discretized (at interval σ) time history of the impulse response of some linear constant dynamical system.

It can happen that the vector s contains the leading coefficients of the expansion in powers of 1/s of a transfer function (the laplace transform of the impulse response). This is, in fact, the originally planned mode of operation for the procedure. In such a case the call to CPC is superfluous. The application for which MICARE was written usually requires the use of CPC, however, and furthermore CPC provides the eigenvalues of ϕ, so no provision was made for avoiding the call to CPC.

II. Mathematical Analysis

1. The B. L. Ho Procedure.

Definition: An infinite matrix is said to have rank r if the maximum rank of any finite submatrix is r.

Proposition 1: Let [c, A, b] be an n^{th} order c.c and c.o stationary system, with impulse response $c\phi(t)b$. Denote $\phi(\delta)$ by ϕ. Let $H = [h_{ij}]$, where

$$h_{ij} = c\phi((i+j-2)\delta)b = c\phi^{i+j-2}b \ ,$$

be an infinite order matrix. Then rank H = n.

Proposition 2: Let [c, A, b] be an n^{th} order c.c. and c.o. stationary system with impulse response $c\phi(t)b = f(t)$. Represent f(t) in its taylor's series expansion

$$\sum_{k=0}^{\infty} \frac{a_k}{k!} t^k \ .$$

Let $H = [h_{ij}]$, where

$$H_{ij} = a_{i+j-2} \ .$$

Then rank H = n.

Proof: Clearly $a_k = cA^k b$, since $a_k = f^{(k)}(0)$. Therefore

$$h_{ij} = cA^{i+j-2}b \ .$$

Also the matrices

$$W_\phi = [b, \phi b, \cdots, \phi^{n-1}b]$$

and

$$W_A =]b, Ab, \cdots , A^{n-1}b]$$

are both nonsingular by complete controllability, as are the comparable observability matrices. These remarks reduce the two propositions to one.

We shall prove proposition 2.

The $n \times m$ matrix $(m \geq n)$

$$W = [b, Ab, A^2b, \cdots]$$

has rank n, as does the $m \times n$ matrix M,

$$M' = [c', A'c', A'^2c', \cdots] \ .$$

Let ν_{ij} denote the elements of MN. Then

$$\nu_{ij} = cA^{i-1}A^{j-1}b = cA^{i+j-2}b \ .$$

That is $MN = H$.

Sylvester's inequality states that

$$\text{rank } M + \text{rank } N - n \leq \text{rank } MN \leq \min(\text{rank } M, \text{rank } N) ,$$

in this case

$$n \leq \text{rank } MN \leq n.$$

Therefore for all $m \geq n$, rank $H = n$.

Remark: Let $F(s) = \mathcal{L}f(t) = \frac{p(s)}{q(s)}$.

Then deg $p <$ deg q. If $F(s)$ is expanded in powers of s^{-1} ,

$$F(s) = \sum_{k=0}^{\infty} \frac{a_k}{s^{k+1}} ,$$

then the a_k are the previously defined taylor coefficients of $f(t)$. This follows, of course, from the fact that

$$\mathcal{L}\, t^k = \frac{k!}{s^k} .$$

Proposition 3: Let $h = [h_{ij}]$ be an (infinite) hankel matrix (i.e. $h_{ij} = v_{i+j-2}$ for some sequence $\{v_k\}$) with n the maximum rank of any submatrix. Then there exists a triple $[c, A, b]$ such that

$$h_{ij} = cA^{i+j-2}b = v_{i+j-2} .$$

Lemma: For such an H, the first n rows $\{R_i\}_1^n$ are linearly independent.

Proof of Lemma: Since every $(n+1)$-rowed submatrix has determinant zero, the first $n+1$ rows are linearly dependent. Therefore there exists a number $r \leq n$ such that $R_1, R_2, \cdots, R_r$ are linearly independent and

$$R_{r+1} = \sum_{k=0}^{r-1} a_k R_{k+1} .$$

From the cyclic character of a hankel matrix, we see that

$$R_{h+q+1} = \sum_{k=0}^{r-1} a_k R_{k+q+1} \quad ,$$

and therefore every row can be expressed in terms of the first r rows. It follows that $r = n$.

Proof of Proposition 3: Let $[a_o, a_1, \cdots, a_{n-1}]$ be the vector defined in the proof of the lemma. Then

$$v_k = cA^k b$$

where $c = (1, 0, \cdots, 0)$

$$b' = (v_o, v_1, \cdots, v_{n-1})$$

and A is the companion-form matrix with last row

$$[a_o, a_1, \cdots, a_{n-1}] \ .$$

Our conclusion from these three propositions is that a hankel matrix has finite rank n iff its sequence is generated by an n^{th} order dynamical system. This is all that is required for the application of the Ho procedure to single-input, single-output systems. However for the multi-input, multi-output case we need one further theorem, the general theorem underlying the Ho procedure, which we give without proof.

Let $[C,A,B]$ be an m-input, p-output, r^{th} order dynamical system and let

$$C\ \phi(k\delta)B = h_k$$

where $h_k = [h_{ijk}]$ is the $p \times m$ matrix which is the impulse response at time $k\delta$.

Define

$$S_{ij} = \begin{bmatrix} h_{ijo} & - & h_{ijn} \\ - & - & - \\ h_{ijn} & - & h_{ij2n} \end{bmatrix}, \qquad \Sigma_{ij} = \begin{bmatrix} h_{ij1} & - & h_{ijn+1} \\ - & - & - \\ h_{ijn+1} & - & h_{ij2n+1} \end{bmatrix}$$

$$S = [S_{ij}] \quad , \quad \text{a} \quad p(n+1) \text{ by } m(n+1) \text{ matrix,}$$

and

$$\Sigma = [\Sigma_{ij}] \quad .$$

In addition, define

$$\tilde{B} = [c_1, c_2, \ldots, c_m]$$

and

$$\tilde{C} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_p \end{bmatrix}$$

where c_i is the $[1+(i-1)\ (n+1)]^{th}$ column of S and r_i is the $[1+(i-1)\ (n+1)]^{th}$ row of S .

Compute nonsingular matrices P and Q such that

$$PSQ = \begin{bmatrix} I_r & 0 \\ 0 & 0 \end{bmatrix} .$$

Then

$$P\Sigma Q = \begin{bmatrix} A & . & \times \\ . & . & . \\ \times & . & \times \end{bmatrix} ,$$

where $\times$ indicates arbitrary elements,

$$\overset{\sim}{PB} = \begin{bmatrix} \hat{G} \\ \cdots \\ \times \end{bmatrix}$$

$$\overset{\sim}{CQ} = [\hat{H} : \times]$$

Then:

Theorem: For sufficiently large $n(n > r)$, the system $[\hat{C},\hat{A},\hat{B}]$ is a minimal order representation of $[C,A,B]$.

2. Computation.

Let H be a hankel matrix of rank n and let S be its first n^{th} order principal submatrix

$$S = \begin{vmatrix} v_o & \cdots & v_{n-1} \\ & \cdots & \\ v_{n-1} & \cdots & v_{2n-2} \end{vmatrix} .$$

By an extension of the lemma, this has rank n.

Compute nonsingular matrices L and R such that

$$LSR = I_n .$$

It follows that $S^{-1} = RL$.

Let

$$S^* = \begin{vmatrix} v_1 & \cdots & v_n \\ & \cdots & \\ v_n & \cdots & v_{2n-1} \end{vmatrix}$$

denote the second n^{th} order principal submatrix of H, and let

$$b' = [v_o, v_1, \cdots, v_{n-1}] .$$

We know that

$$S^* = AS$$

where A is the matrix defined in proving proposition 3.

Compute

$$c^* = b'R ,$$

$$b^* = Lb ,$$

and $A^* = LS^*R = LASR$. Then

$$c^*b^* = b'RLb = (1, 0, \cdots, 0)b = v_o$$

and

$$c^*A^{*k}b^* = b'R(LASR)^kLb = b'RL(ASRL)^kb = cA^kb .$$

To provide additional smoothing, we compute with a rectangular matrix having more rows than columns, with a maximum of 15 columns and 100 rows,

$$S = \begin{bmatrix} v_o & \cdots & v_{m-1} \\ & \cdots & \\ v_{r-1} & \cdots & v_{m+r-2} \end{bmatrix} .$$

Here we have $r \gg n = \text{rank } H' m \geq n$.

We find matrices L and R of rank n such that

$$LSR = I_n$$

and

$$SR = L' \quad .$$

Lemma: $SRLS = S$.

Proof: Since $SR = L'$ and rank L = rank S, L is nonsingular on range S, therefore the fact that

$$L(SRLS - S) = LS - LS = 0 \ ,$$

implies that $SRLS - S = 0$.

Let

$$S^* = \begin{bmatrix} v_1 & \cdots & v_m \\ & \cdots & \\ v_r & \cdots & v_{m+r-1} \end{bmatrix} \quad .$$

We can define an m by m matrix $\hat{A}$ such that $S^* = \hat{A}S$. In the 3×3 case, if

$$A = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ a & b & c \end{bmatrix}$$

$$\hat{A} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & \cdot \\ 0 & 0 & 1 & 0 & 0 & \cdot \\ a & b & c & 0 & 0 & \cdot \\ 0 & a & b & c & 0 & \cdot \\ 0 & 0 & a & b & c & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \end{bmatrix}$$

to the size required.

The important thing is that b is the first column of S and $\hat{A}^k b$ is b_{k+1}, the $(k+1)^{st}$ column of S (or the first m rows of the $(k+1)^{st}$ column of H if $k \geq m$).

We compute

$$c^* = b'R ,$$

$$b^* = Lb ,$$

$$A^* = LS^*R = L\hat{A}SR .$$

Then

$$c^*b^* - b'RLb = v_o$$

by the lemma.

$$c^*A^*b^* = b'RL\hat{A}SRLb - b'RL\hat{A}b$$

by the lemma.

But $\hat{A}b$ is the second column of S so $b'RL\hat{A}b = v_1$, again by using the lemma.

In general

$$c^*A^{*k}b^* - b'R(L\hat{A}SR)^kLb = b'RL(\hat{A}SRL)^kb .$$

By induction we can show that

$$(\hat{A}SRL)^k b = b_{k+1} .$$

Since $b_{k+1} \in$ range (S), it follows by the lemma that

$$b'RLb_{k+1} = v_k .$$

Remark: Notice that RL need not be the generalized inverse of S but must satisfy only

$$SRLS = S \ .$$

In the multi-input, multi-output case we change the construction of the hankel matrix from that described in section 1 by rearranging the columns. The rearrangement puts first in S the columns $[c_1, c_2, \cdots, c_m]$ which we used to define $\tilde{B}$ in 2. The rest of S is then filled out with the successive matrices $\tilde{A}\tilde{B}$, $\tilde{A}^2\tilde{B}$, etc.

The routine we are using to compute L and R has a bias toward accepting the early columns of a matrix as linearly independent and therefore this rearrangement was performed in order to insure that inputs other than the first will have good numerical significance.

3. Mechanization.

Starting with an S having NST columns, we find matrices T_L, T_R such that

$$T_L \ S \ T_R = I$$

where I is an n-dimensional identity, T_L and T_R are saved.

Increasing the dimension of H by one we replace T_L and T_R by their new values if the rank increases.

If the rank is unchanged, either because of the constraint IDERK or because the rank is the same within the tolerance TOLI, we use the T_L and T_R from the previous dimension KM1 as follows.

The matrix S* of dimension KM1 is formed

$$S^* = \begin{bmatrix} s_2 & s_3 & s_4 & \cdot \\ s_3 & s_4 & s_5 & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}$$

Then the system matrix is $\phi^* = T_L\ S^*\ T_R$, the output vector is $c^* = [s_1, \cdots, s_{KM1}]T_R$, and the input vector is

$$\gamma^* = T_L \begin{bmatrix} s_1 \\ s_2 \\ \cdot \\ \cdot \\ \cdot \\ s_{KM1} \end{bmatrix}$$

The impulse response $\hat{s}_k = c^*\ \phi^{*k-1}\ \gamma^*$ of this system is compared with s_k, the system is put in companion form, and the logarithm system computed.

If a reasonable approximation between S and $\hat{S}$ was found, we RETURN. If a term was too much in error, the dimension of S is increased to include that term in the next system. This proceeds until a good fit is obtained or the S vector is exhausted.

APPENDIX III

Program CPC

SUBROUTINE CPC (S, IRANK, B, C, DT)

Purpose: We are given the $n \times n$ companion form matrix ϕ, vectors G and H, and a time increment δ. We wish to find an $n \times n$ companion form matrix A and vectors B and C ($C = \pm 1, 0, 0, \ldots 01$) such that

$$Ce^{k\delta A}B = H\phi^{k}G, \quad k = 0,1,\ldots .$$

Basically we wish to find the logarithm of ϕ.

Restrictions and Commentary:

1) Naturally ϕ must be nonsingular.

2) ϕ cannot have repeated eigenvalues. In practice this is not a very serious restriction. Numerical difficulties may occur when roots are close to each other.

3) Early in the program, eigenvalues $\lambda = x + iy$ are assumed to be real and positive if they satisfy

$$\frac{|y|}{10^{-7} + |x|} < 10^{-7}$$

Theoretically this is a vulnerable point. If there is a complex pair of ϕ with small imaginary part, trouble can occur. However, this is essentially covered by the restriction that roots must be distinct. Perhaps more important, a complex pair in F can, for proper values of the time increment, give rise to a coincident pair of negative eigenvalues of ϕ. However, we do not expect this to occur because good engineering practice will dictate that the time increment used to generate ϕ will be selected less than half the natural period.

Besides which the condition is highly improbable under any circumstances.

4) This program, because of the application which evoked it, assumes that the pair $[H, \phi]$ is completely observable. This is clear from the output form of C and A.

Procedure: Since ϕ is given in companion form, the characteristic polynomial is immediately available. This is factored to obtain the eigenvalues of ϕ. If the eigenvalue $\lambda = x + iy$ satisfies

$$\frac{|y|}{10^{-7} + |x|} < 10^{-7}$$

the eigenvalue is taken as real and positive, otherwise as complex.

We set up a complex n-vector with the complex roots first and the real roots last.

The eigenvalues of ϕ are printed. The number of complex roots is printed.

The generalized Vandermonde matrix T is constructed which transforms ϕ to its real diagonal form, R.

T is inverted to form T^{-1}.

HT and ϕT are formed.

$T^{-1}G$ and $T^{-1}\phi T$ are formed.

$T^{-1}\phi T$ is printed. The computation and subsequent printout of

$T^{-1}\phi T$ is done purely as a numerical check since $T^{-1}\phi T$ will be assumed to have the correct real diagonal form R and its computed value destroyed after printing.

$M = \log R$ is constructed and printed. Following this, the matrix

$$S = \begin{bmatrix} HT \\ HTM \\ - \\ - \\ - \\ HTM^{n-1} \end{bmatrix}$$

is formed, and finally the desired matrices

$$C = HTS^{-1}$$

$$A = SMS^{-1}$$

$$B = ST^{-1}G$$

are printed.

Mathematical analysis:

1) Real diagonal form and generalized Vandermonde:

If a matrix ϕ has only real eigenvalues, its real diagonal form Λ is its diagonal form and the matrix T transforming to Λ is the Vandermonde

$$T^{-1} T = \Lambda.$$

Where $t_{ij} = \lambda_j^{i-1}$.

If there is a single complex pair $a \pm bi$ then we take

$$T = \begin{bmatrix} 1 & 0 \\ a & b \end{bmatrix} \qquad T^{-1} = \begin{bmatrix} 1 & 0 \\ -\frac{a}{b} & \frac{1}{b} \end{bmatrix}$$

$$\phi = \begin{bmatrix} 0 & 1 \\ -(a^2+b^2) & 2a \end{bmatrix}$$

$$T^{-1}\phi T = \begin{bmatrix} a & b \\ -b & a \end{bmatrix}$$

We call this the real diagonal form for this ϕ. In general, if there are r complex roots, the real diagonal form for ϕ is the direct sum of r such 2×2 matrices and an $(n - r)$-dimensional diagonal matrix. The jth column of the generalized Vandermonde T corresponding to a real root λ_j is $t_{ij} = \lambda_j^{i-1}$. The columns, say 1 and 2, corresponding to the pair $\lambda_1 = a + ib$ and $\lambda_2 = a - ib$ are

$$t_{i1} = \mathrm{Re}(\lambda_1^{i-1})$$

$$t_{i2} = \mathrm{Im}(\lambda_1^{i-1}) \ .$$

The first such column starts

$$1,\ a,\ a^2 - b^2,\ a^3 - 3ab^2,\ \ldots$$

the second such column starts

$$0,\ b,\ 2ab,\ 3a^2b - b^3,\ \ldots$$

2) Logarithm of the real diagonal form.

Let R denote the real diagonal form.

The logarithm of the diagonal part of R is very simple being the diagonal matrix M whose elements are the logarithms of the (positive real) diagonal elements of R.

The rest of R is the direct sum of 2 x 2 matrices of the form

$$\begin{bmatrix} a & b \\ -b & a \end{bmatrix} .$$

The logarithm of this matrix is

$$\begin{bmatrix} \log(a^2 + b^2) & \tan^{-1}\frac{b}{a} \\ -\tan^{-1}\frac{b}{a} & \log(a^2 + b^2) \end{bmatrix} .$$

The nondiagonal part of M is the direct sum of such 2 x 2 matrices.

As is well know, the logarithm is not uniquely defined. Naturally we take the smallest value of the imaginary part which will give the correct

exponential. Our justification for this is again the assumption that the δ used to generate ϕ was smaller than half the smallest natural period appearing in the spectrum of A.

3) Companion form.

An nth order matrix A is said to be in companion form if

$$a_{i,i+1} = 1 \quad ,$$

the characteristic polynomial of A is

$$x^n + \sum_{j=o}^{n-1} a_{n,j+1} x^j \quad ,$$

and all other a_{ij}, besides the last row and the first upper diagonal, are zero.

It is easy to show that if the matrix

$$S = \begin{bmatrix} H \\ H \\ \cdot \\ \cdot \\ \cdot \\ H\phi^{n-1} \end{bmatrix}$$

is nonsingular, i.e., if $[H, \phi]$ is completely observable, then

$$HS^{-1} = [1, 0, \cdots, 0]$$

and $S\phi S^{-1}$ is in companion form.